Beyond Value-Function Gaps: Improved Instance-Dependent Regret Bounds for Episod
ic Reinforcement Learning
Christoph Dann, Teodor Vanislavov Marinov, Mehryar Mohri, Julian Zimmert
We provide improved gap-dependent regret bounds for reinforcement learning in fi
nite episodic Markov decision processes. Compared to prior work, our bounds depe
nd on alternative definitions of gaps. These definitions are based on the insigh
t that, in order to achieve a favorable regret, an algorithm does not need to le
arn how to behave optimally in states that are not reached by an optimal policy.
 We prove tighter upper regret bounds for optimistic algorithms and accompany th
em with new information-theoretic lower bounds for a large class of MDPs. Our re
sults show that optimistic algorithms can not achieve the information-theoretic
lower bounds even in deterministic MDPs unless there is a unique optimal policy.
****
Learning One Representation to Optimize All Rewards
Ahmed Touati, Yann Ollivier
We introduce the forward-backward (FB) representation of the dynamics of a rewar
d-free Markov decision process. It provides explicit near-optimal policies for a
ny reward specified a posteriori. During an unsupervised phase, we use reward-fr
ee interactions with the environment to learn two representations via off-the-sh
elf deep learning methods and temporal difference (TD) learning. In the test pha
se, a reward representation is estimated either from reward observations or an e
xplicit reward description (e.g., a target state). The optimal policy for thatre
ward is directly obtained from these representations, with no planning. We assum
e access to an exploration scheme or replay buffer for the first phase.The corre
sponding unsupervised loss is well-principled: if training is perfect, the polic
ies obtained are provably optimal for any reward function. With imperfect train
ing, the sub-optimality is proportional to the unsupervised approximation error.
 The FB representation learns long-range relationships between states and action
s, via a predictive occupancy map, without having to synthesize states as in mod
el-based approaches.This is a step towards learning controllable agents in arbit
rary black-box stochastic environments. This approach compares well to goal-orie
nted RL algorithms on discrete and continuous mazes, pixel-based MsPacman, and t
he FetchReach virtual robot arm. We also illustrate how the agent can immediatel
y adapt to new tasks beyond goal-oriented RL.
****
Matrix factorisation and the interpretation of geodesic distance
Nick Whiteley, Annie Gray, Patrick Rubin-Delanchy
Given a graph or similarity matrix, we consider the problem of recovering a noti
on of true distance between the nodes, and so their true positions. We show that
 this can be accomplished in two steps: matrix factorisation, followed by nonlin
ear dimension reduction. This combination is effective because the point cloud o
btained in the first step lives close to a manifold in which latent distance is
encoded as geodesic distance. Hence, a nonlinear dimension reduction tool, appro
ximating geodesic distance, can recover the latent positions, up to a simple tra
nsformation. We give a detailed account of the case where spectral embedding is
used, followed by Isomap, and provide encouraging experimental evidence for othe
r combinations of techniques.
****
UniDoc: Unified Pretraining Framework for Document Understanding
Jiuxiang Gu, Jason Kuen, Vlad I Morariu, Handong Zhao, Rajiv Jain, Nikolaos Barm
palios, Ani Nenkova, Tong Sun
Document intelligence automates the extraction of information from documents and
 supports many business applications. Recent self-supervised learning methods on
 large-scale unlabeled document datasets have opened up promising directions tow
ards reducing annotation efforts by training models with self-supervised objecti
ves. However, most of the existing document pretraining methods are still langua
ge-dominated. We present UDoc, a new unified pretraining framework for document
understanding. UDoc is designed to support most document understanding tasks, ex
tending the Transformer to take multimodal embeddings as input. Each input eleme
nt is composed of words and visual features from a semantic region of the input

document image. An important feature of UDoc is that it learns a generic represe ntation by making use of three self-supervised losses, encouraging the represent ation to model sentences, learn similarities, and align modalities. Extensive em pirical analysis demonstrates that the pretraining procedure learns better joint representations and leads to improvements in downstream tasks.
****
Finding Discriminative Filters for Specific Degradations in Blind Super-Resoluti on
Liangbin Xie, Xintao Wang, Chao Dong, Zhongang Qi, Ying Shan
Recent blind super-resolution (SR) methods typically consist of two branches, on e for degradation prediction and the other for conditional restoration. However, our experiments show that a one-branch network can achieve comparable performan ce to the two-branch scheme. Then we wonder: how can one-branch networks automat ically learn to distinguish degradations? To find the answer, we propose a new d iagnostic tool -- Filter Attribution method based on Integral Gradient (FAIG). U nlike previous integral gradient methods, our FAIG aims at finding the most disc riminative filters instead of input pixels/features for degradation removal in b lind SR networks. With the discovered filters, we further develop a simple yet e ffective method to predict the degradation of an input image. Based on FAIG, we show that, in one-branch blind SR networks, 1) we could find a very small number of (1%) discriminative filters for each specific degradation; 2) The weights, l ocations and connections of the discovered filters are all important to determin e the specific network function. 3) The task of degradation prediction can be im plicitly realized by these discriminative filters without explicit supervised le arning. Our findings can not only help us better understand network behaviors in side one-branch blind SR networks, but also provide guidance on designing more e fficient architectures and diagnosing networks for blind SR.
****
Counterfactual Explanations Can Be Manipulated
Dylan Slack, Anna Hilgard, Himabindu Lakkaraju, Sameer Singh
Counterfactual explanations are emerging as an attractive option for providing r ecourse to individuals adversely impacted by algorithmic decisions. As they are deployed in critical applications (e.g. law enforcement, financial lending), it becomes important to ensure that we clearly understand the vulnerabilties of th ese methods and find ways to address them. However, there is little understandin g of the vulnerabilities and shortcomings of counterfactual explanations. In thi s work, we introduce the first framework that describes the vulnerabilities of c ounterfactual explanations and shows how they can be manipulated. More specifica lly, we show counterfactual explanations may converge to drastically different c ounterfactuals under a small perturbation indicating they are not robust. Lever aging this insight, we introduce a novel objective to train seemingly fair model s where counterfactual explanations find much lower cost recourse under a slight perturbation. We describe how these models can unfairly provide low-cost recou rse for specific subgroups in the data while appearing fair to auditors. We perf orm experiments on loan and violent crime prediction data sets where certain sub groups achieve up to 20x lower cost recourse under the perturbation. These resul ts raise concerns regarding the dependability of current counterfactual explanat ion techniques, which we hope will inspire investigations in robust counterfactu al explanations.
****
From Canonical Correlation Analysis to Self-supervised Graph Neural Networks
Hengrui Zhang, Qitian Wu, Junchi Yan, David Wipf, Philip S Yu
We introduce a conceptually simple yet effective model for self-supervised repre sentation learning with graph data. It follows the previous methods that generat e two views of an input graph through data augmentation. However, unlike contras tive methods that focus on instance-level discrimination, we optimize an innovat ive feature-level objective inspired by classical Canonical Correlation Analysis . Compared with other works, our approach requires none of the parameterized mut ual information estimator, additional projector, asymmetric structures, and most importantly, negative samples which can be costly. We show that the new objecti

ve essentially 1) aims at discarding augmentation-variant information by learnin
g invariant representations, and 2) can prevent degenerated solutions by decorre
lating features in different dimensions. Our theoretical analysis further provid
es an understanding for the new objective which can be equivalently seen as an i
nstantiation of the Information Bottleneck Principle under the self-supervised s
etting. Despite its simplicity, our method performs competitively on seven publi
c graph datasets.
****
BAST: Bayesian Additive Regression Spanning Trees for Complex Constrained Domain
Zhao Tang Luo, Huiyan Sang, Bani Mallick
Nonparametric regression on complex domains has been a challenging task as most
existing methods, such as ensemble models based on binary decision trees, are no
t designed to account for intrinsic geometries and domain boundaries. This artic
le proposes a Bayesian additive regression spanning trees (BAST) model for nonpa
rametric regression on manifolds, with an emphasis on complex constrained domain
s or irregularly shaped spaces embedded in Euclidean spaces. Our model is built
upon a random spanning tree manifold partition model as each weak learner, which
 is capable of capturing any irregularly shaped spatially contiguous partitions
while respecting intrinsic geometries and domain boundary constraints. Utilizing
 many nice properties of spanning tree structures, we design an efficient Bayesi
an inference algorithm. Equipped with a soft prediction scheme, BAST is demonstr
ated to significantly outperform other competing methods in simulation experimen
ts and in an application to the chlorophyll data in Aral Sea, due to its strong
local adaptivity to different levels of smoothness.
****
Hyperbolic Busemann Learning with Ideal Prototypes
Mina Ghadimi Atigh, Martin Keller-Ressel, Pascal Mettes
Hyperbolic space has become a popular choice of manifold for representation lear
ning of various datatypes from tree-like structures and text to graphs. Building
 on the success of deep learning with prototypes in Euclidean and hyperspherical
 spaces, a few recent works have proposed hyperbolic prototypes for classificati
on. Such approaches enable effective learning in low-dimensional output spaces a
nd can exploit hierarchical relations amongst classes, but require privileged in
formation about class labels to position the hyperbolic prototypes. In this work
, we propose Hyperbolic Busemann Learning. The main idea behind our approach is
to position prototypes on the ideal boundary of the Poincar\'{e} ball, which doe
s not require prior label knowledge. To be able to compute proximities to ideal
prototypes, we introduce the penalised Busemann loss. We provide theory supporti
ng the use of ideal prototypes and the proposed loss by proving its equivalence
to logistic regression in the one-dimensional case. Empirically, we show that ou
r approach provides a natural interpretation of classification confidence, while
 outperforming recent hyperspherical and hyperbolic prototype approaches.
****
Backward-Compatible Prediction Updates: A Probabilistic Approach
Frederik Träuble, Julius von Kügelgen, Matthäus Kleindessner, Francesco Locatell
o, Bernhard Schölkopf, Peter Gehler
When machine learning systems meet real world applications, accuracy is only one
 of several requirements. In this paper, we assay a complementary perspective or
iginating from the increasing availability of pre-trained and regularly improvin
g state-of-the-art models. While new improved models develop at a fast pace, dow
nstream tasks vary more slowly or stay constant. Assume that we have a large unl
abelled data set for which we want to maintain accurate predictions. Whenever a
new and presumably better ML models becomes available, we encounter two problems
: (i) given a limited budget, which data points should be re-evaluated using the
 new model?; and (ii) if the new predictions differ from the current ones, shoul
d we update? Problem (i) is about compute cost, which matters for very large dat
a sets and models. Problem (ii) is about maintaining consistency of the predicti
ons, which can be highly relevant for downstream applications; our demand is to
avoid negative flips, i.e., changing correct to incorrect predictions. In this p
aper, we formalize the Prediction Update Problem and present an efficient probab

ilistic approach as answer to the above questions. In extensive experiments on standard classification benchmark data sets, we show that our method outperforms alternative strategies along key metrics for backward-compatible prediction updates.
****
Truncated Marginal Neural Ratio Estimation
Benjamin K Miller, Alex Cole, Patrick Forré, Gilles Louppe, Christoph Weniger
Parametric stochastic simulators are ubiquitous in science, often featuring high-dimensional input parameters and/or an intractable likelihood. Performing Bayesian parameter inference in this context can be challenging. We present a neural simulation-based inference algorithm which simultaneously offers simulation efficiency and fast empirical posterior testability, which is unique among modern algorithms. Our approach is simulation efficient by simultaneously estimating low-dimensional marginal posteriors instead of the joint posterior and by proposing simulations targeted to an observation of interest via a prior suitably truncated by an indicator function. Furthermore, by estimating a locally amortized posterior our algorithm enables efficient empirical tests of the robustness of the inference results. Since scientists cannot access the ground truth, these tests are necessary for trusting inference in real-world applications. We perform experiments on a marginalized version of the simulation-based inference benchmark and two complex and narrow posteriors, highlighting the simulator efficiency of our algorithm as well as the quality of the estimated marginal posteriors.
****
ReAct: Out-of-distribution Detection With Rectified Activations
Yiyou Sun, Chuan Guo, Yixuan Li
Out-of-distribution (OOD) detection has received much attention lately due to its practical importance in enhancing the safe deployment of neural networks. One of the primary challenges is that models often produce highly confident predictions on OOD data, which undermines the driving principle in OOD detection that the model should only be confident about in-distribution samples. In this work, we propose ReAct—a simple and effective technique for reducing model overconfidence on OOD data. Our method is motivated by novel analysis on internal activations of neural networks, which displays highly distinctive signature patterns for OOD distributions. Our method can generalize effectively to different network architectures and different OOD detection scores. We empirically demonstrate that ReAct achieves competitive detection performance on a comprehensive suite of benchmark datasets, and give theoretical explication for our method’s efficacy. On the ImageNet benchmark, ReAct reduces the false positive rate (FPR95) by 25.05% compared to the previous best method.
****
Non-local Latent Relation Distillation for Self-Adaptive 3D Human Pose Estimation
Jogendra Nath Kundu, Siddharth Seth, Anirudh Jamkhandi, Pradyumna YM, Varun Jampani, Anirban Chakraborty, Venkatesh Babu R
Available 3D human pose estimation approaches leverage different forms of strong (2D/3D pose) or weak (multi-view or depth) paired supervision. Barring synthetic or in-studio domains, acquiring such supervision for each new target environment is highly inconvenient. To this end, we cast 3D pose learning as a self-supervised adaptation problem that aims to transfer the task knowledge from a labeled source domain to a completely unpaired target. We propose to infer image-to-pose via two explicit mappings viz. image-to-latent and latent-to-pose where the latter is a pre-learned decoder obtained from a prior-enforcing generative adversarial auto-encoder. Next, we introduce relation distillation as a means to align the unpaired cross-modal samples i.e., the unpaired target videos and unpaired 3D pose sequences. To this end, we propose a new set of non-local relations in order to characterize long-range latent pose interactions, unlike general contrastive relations where positive couplings are limited to a local neighborhood structure. Further, we provide an objective way to quantify non-localness in order to select the most effective relation set. We evaluate different self-adaptation settings and demonstrate state-of-the-art 3D human pose estimation performance on

standard benchmarks.
****
Fast Training of Neural Lumigraph Representations using Meta Learning
Alexander Bergman, Petr Kellnhofer, Gordon Wetzstein
Novel view synthesis is a long-standing problem in machine learning and computer vision. Significant progress has recently been made in developing neural scene representations and rendering techniques that synthesize photorealistic images from arbitrary views. These representations, however, are extremely slow to train and often also slow to render. Inspired by neural variants of image-based rendering, we develop a new neural rendering approach with the goal of quickly learning a high-quality representation which can also be rendered in real-time. Our approach, MetaNLR++, accomplishes this by using a unique combination of a neural shape representation and 2D CNN-based image feature extraction, aggregation, and re-projection. To push representation convergence times down to minutes, we leverage meta learning to learn neural shape and image feature priors which accelerate training. The optimized shape and image features can then be extracted using traditional graphics techniques and rendered in real time. We show that MetaNLR++ achieves similar or better novel view synthesis results in a fraction of the time that competing methods require.
****
Analytical Study of Momentum-Based Acceleration Methods in Paradigmatic High-Dimensional Non-Convex Problems
Stefano Sarao Mannelli, Pierfrancesco Urbani
The optimization step in many machine learning problems rarely relies on vanilla gradient descent but it is common practice to use momentum-based accelerated methods. Despite these algorithms being widely applied to arbitrary loss functions, their behaviour in generically non-convex, high dimensional landscapes is poorly understood.In this work, we use dynamical mean field theory techniques to describe analytically the average dynamics of these methods in a prototypical non-convex model: the (spiked) matrix-tensor model. We derive a closed set of equations that describe the behaviour of heavy-ball momentum and Nesterov acceleration in the infinite dimensional limit. By numerical integration of these equations, we observe that these methods speed up the dynamics but do not improve the algorithmic threshold with respect to gradient descent in the spiked model.
****
Multimodal Few-Shot Learning with Frozen Language Models
Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, Felix Hill
When trained at sufficient scale, auto-regressive language models exhibit the notable ability to learn a new language task after being prompted with just a few examples. Here, we present a simple, yet effective, approach for transferring this few-shot learning ability to a multimodal setting (vision and language). Using aligned image and caption data, we train a vision encoder to represent each image as a sequence of continuous embeddings, such that a pre-trained, frozen language model presented with this prefix generates the appropriate caption. The resulting system is a multimodal few-shot learner, with the surprising ability to learn a variety of new tasks when conditioned on examples, represented as a sequence of any number of interleaved image and text embeddings. We demonstrate that it can rapidly learn words for new objects and novel visual categories, do visual question-answering with only a handful of examples, and make use of outside knowledge, by measuring a single model on a variety of established and new benchmarks.
****
Approximating the Permanent with Deep Rejection Sampling
Juha Harviainen, Antti Röyskö, Mikko Koivisto
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

Revisiting Model Stitching to Compare Neural Representations
Yamini Bansal, Preetum Nakkiran, Boaz Barak
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
AugMax: Adversarial Composition of Random Augmentations for Robust Training
Haotao Wang, Chaowei Xiao, Jean Kossaifi, Zhiding Yu, Anima Anandkumar, Zhangyang Wang
Data augmentation is a simple yet effective way to improve the robustness of deep neural networks (DNNs). Diversity and hardness are two complementary dimensions of data augmentation to achieve robustness. For example, AugMix explores random compositions of a diverse set of augmentations to enhance broader coverage, while adversarial training generates adversarially hard samples to spot the weakness. Motivated by this, we propose a data augmentation framework, termed AugMax, to unify the two aspects of diversity and hardness. AugMax first randomly samples multiple augmentation operators and then learns an adversarial mixture of the selected operators. Being a stronger form of data augmentation, AugMax leads to a significantly augmented input distribution which makes model training more challenging. To solve this problem, we further design a disentangled normalization module, termed DuBIN (Dual-Batch-and-Instance Normalization), that disentangles the instance-wise feature heterogeneity arising from AugMax. Experiments show that AugMax-DuBIN leads to significantly improved out-of-distribution robustness, outperforming prior arts by 3.03%, 3.49%, 1.82% and 0.71% on CIFAR10-C, CIFAR100-C, Tiny ImageNet-C and ImageNet-C. Codes and pretrained models are available: https://github.com/VITA-Group/AugMax.
****
Habitat 2.0: Training Home Assistants to Rearrange their Habitat
Andrew Szot, Alexander Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Singh Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimír Vondruš, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, Manolis Savva, Dhruv Batra
We introduce Habitat 2.0 (H2.0), a simulation platform for training virtual robots in interactive 3D environments and complex physics-enabled scenarios. We make comprehensive contributions to all levels of the embodied AI stack – data, simulation, and benchmark tasks. Specifically, we present: (i) ReplicaCAD: an artist-authored, annotated, reconfigurable 3D dataset of apartments (matching real spaces) with articulated objects (e.g. cabinets and drawers that can open/close); (ii) H2.0: a high-performance physics-enabled 3D simulator with speeds exceeding 25,000 simulation steps per second (850x real-time) on an 8-GPU node, representing 100x speed-ups over prior work; and, (iii) Home Assistant Benchmark (HAB): a suite of common tasks for assistive robots (tidy the house, stock groceries, set the table) that test a range of mobile manipulation capabilities. These large-scale engineering contributions allow us to systematically compare deep reinforcement learning (RL) at scale and classical sense-plan-act (SPA) pipelines in long-horizon structured tasks, with an emphasis on generalization to new objects, receptacles, and layouts. We find that (1) flat RL policies struggle on HAB compared to hierarchical ones; (2) a hierarchy with independent skills suffers from ‘hand-off problems’, and (3) SPA pipelines are more brittle than RL policies.
****
Time Discretization-Invariant Safe Action Repetition for Policy Gradient Methods
Seohong Park, Jaekyeom Kim, Gunhee 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.
****
Meta-Learning Reliable Priors in the Function Space

Jonas Rothfuss, Dominique Heyn, jinfan Chen, Andreas Krause
Meta-Learning promises to enable more data-efficient inference by harnessing pre vious experience from related learning tasks. While existing meta-learning metho ds help us to improve the accuracy of our predictions in face of data scarcity, they fail to supply reliable uncertainty estimates, often being grossly overconf ident in their predictions. Addressing these shortcomings, we introduce a novel meta-learning framework, called F-PACOH, that treats meta-learned priors as stoc hastic processes and performs meta-level regularization directly in the function space. This allows us to directly steer the probabilistic predictions of the me ta-learner towards high epistemic uncertainty in regions of insufficient meta-tr aining data and, thus, obtain well-calibrated uncertainty estimates. Finally, we showcase how our approach can be integrated with sequential decision making, wh ere reliable uncertainty quantification is imperative. In our benchmark study on meta-learning for Bayesian Optimization (BO), F-PACOH significantly outperforms all other meta-learners and standard baselines. Even in a challenging lifelong BO setting, where optimization tasks arrive one at a time and the meta-learner needs to build up informative prior knowledge incrementally, our proposed method demonstrates strong positive transfer.
****
VoiceMixer: Adversarial Voice Style Mixup
Sang-Hoon Lee, Ji-Hoon Kim, Hyunseung Chung, Seong-Whan Lee
Although recent advances in voice conversion have shown significant improvement, there still remains a gap between the converted voice and target voice. A key f actor that maintains this gap is the insufficient decomposition of content and v oice style from the source speech. This insufficiency leads to the converted spe ech containing source speech style or losing source speech content. In this pape r, we present VoiceMixer which can effectively decompose and transfer voice styl e through a novel information bottleneck and adversarial feedback. With self-sup ervised representation learning, the proposed information bottleneck can decompo se the content and style with only a small loss of content information. Also, fo r adversarial feedback of each information, the discriminator is decomposed into content and style discriminator with self-supervision, which enable our model t o achieve better generalization to the voice style of the converted speech. The experimental results show the superiority of our model in disentanglement and tr ansfer performance, and improve audio quality by preserving content information.
****
Predicting What You Already Know Helps: Provable Self-Supervised Learning
Jason D. Lee, Qi Lei, Nikunj Saunshi, JIACHENG ZHUO
Self-supervised representation learning solves auxiliary prediction tasks (known as pretext tasks), that do not require labeled data, to learn semantic represen tations. These pretext tasks are created solely using the input features, such a s predicting a missing image patch, recovering the color channels of an image fr om context, or predicting missing words, yet predicting this \textit{known} info rmation helps in learning representations effective for downstream prediction ta sks. This paper posits a mechanism based on approximate conditional independence to formalize how solving certain pretext tasks can learn representations that p rovably decrease the sample complexity of downstream supervised tasks. Formally, we quantify how the approximate independence between the components of the pret ext task (conditional on the label and latent variables) allows us to learn repr esentations that can solve the downstream task with drastically reduced sample c omplexity by just training a linear layer on top of the learned representation.
****
Oracle Complexity in Nonsmooth Nonconvex Optimization
Guy Kornowski, 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.
****
CentripetalText: An Efficient Text Instance Representation for Scene Text Detect

ion
Tao Sheng, Jie Chen, Zhouhui Lian
Scene text detection remains a grand challenge due to the variation in text curvatures, orientations, and aspect ratios. One of the hardest problems in this task is how to represent text instances of arbitrary shapes. Although many methods have been proposed to model irregular texts in a flexible manner, most of them lose simplicity and robustness. Their complicated post-processings and the regression under Dirac delta distribution undermine the detection performance and the generalization ability. In this paper, we propose an efficient text instance representation named CentripetalText (CT), which decomposes text instances into the combination of text kernels and centripetal shifts. Specifically, we utilize the centripetal shifts to implement pixel aggregation, guiding the external text pixels to the internal text kernels. The relaxation operation is integrated into the dense regression for centripetal shifts, allowing the correct prediction in a range instead of a specific value. The convenient reconstruction of text contours and the tolerance of prediction errors in our method guarantee the high detection accuracy and the fast inference speed, respectively. Besides, we shrink our text detector into a proposal generation module, namely CentripetalText Proposal Network (CPN), replacing Segmentation Proposal Network (SPN) in Mask TextSpotter v3 and producing more accurate proposals. To validate the effectiveness of our method, we conduct experiments on several commonly used scene text benchmarks, including both curved and multi-oriented text datasets. For the task of scene text detection, our approach achieves superior or competitive performance compared to other existing methods, e.g., F-measure of 86.3% at 40.0 FPS on Total-Text, F-measure of 86.1% at 34.8 FPS on MSRA-TD500, etc. For the task of end-to-end scene text recognition, our method outperforms Mask TextSpotter v3 by 1.1% in F-measure on Total-Text.
****
Learning to Select Exogenous Events for Marked Temporal Point Process
Ping Zhang, Rishabh Iyer, Ashish Tendulkar, Gaurav Aggarwal, Abir De
Marked temporal point processes (MTPPs) have emerged as a powerful modelingtool for a wide variety of applications which are characterized using discreteevents localized in continuous time. In this context, the events are of two typesendogenous events which occur due to the influence of the previous events andexogenous events which occur due to the effect of the externalities. However, inpractice, the events do not come with endogenous or exogenous labels. To thisend, our goal in this paper is to identify the set of exogenous events from a set ofunlabelled events. To do so, we first formulate the parameter estimation problemin conjunction with exogenous event set selection problem and show that thisproblem is NP hard. Next, we prove that the underlying objective is a monotoneand \alpha-submodular set function, with respect to the candidate set of exogenousevents. Such a characterization subsequently allows us to use a stochastic greedyalgorithm which was originally proposed in~\cite{greedy}for submodular maximization.However, we show that it also admits an approximation guarantee for maximizing\alpha-submodular set function, even when the learning algorithm provides an imperfectestimates of the trained parameters. Finally, our experiments with synthetic andreal data show that our method performs better than the existing approaches builtupon superposition of endogenous and exogenous MTPPs.
****
DRIVE: One-bit Distributed Mean Estimation
Shay Vargaftik, Ran Ben-Basat, Amit Portnoy, Gal Mendelson, Yaniv Ben-Itzhak, Michael Mitzenmacher
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 Space Partitions for Path Planning
Kevin Yang, Tianjun Zhang, Chris Cummins, Brandon Cui, Benoit Steiner, Linnan Wang, Joseph E. Gonzalez, Dan Klein, Yuandong Tian

Path planning, the problem of efficiently discovering high-reward trajectories, often requires optimizing a high-dimensional and multimodal reward function. Popular approaches like CEM and CMA-ES greedily focus on promising regions of the search space and may get trapped in local maxima. DOO and VOOT balance exploration and exploitation, but use space partitioning strategies independent of the reward function to be optimized. Recently, LaMCTS empirically learns to partition the search space in a reward-sensitive manner for black-box optimization. In this paper, we develop a novel formal regret analysis for when and why such an adaptive region partitioning scheme works. We also propose a new path planning method LaP3 which improves the function value estimation within each sub-region, and uses a latent representation of the search space. Empirically, LaP3 outperforms existing path planning methods in 2D navigation tasks, especially in the presence of difficult-to-escape local optima, and shows benefits when plugged into the planning components of model-based RL such as PETS. These gains transfer to highly multimodal real-world tasks, where we outperform strong baselines in compiler phase ordering by up to 39% on average across 9 tasks, and in molecular design by up to 0.4 on properties on a 0-1 scale. Code is available at https://github.com/yangkevin2/neurips2021-lap3.
****
Progressive Feature Interaction Search for Deep Sparse Network
Chen Gao, Yinfeng Li, Quanming Yao, Depeng Jin, Yong Li
Deep sparse networks (DSNs), of which the crux is exploring the high-order feature interactions, have become the state-of-the-art on the prediction task with high-sparsity features. However, these models suffer from low computation efficiency, including large model size and slow model inference, which largely limits these models' application value. In this work, we approach this problem with neural architecture search by automatically searching the critical component in DSNs, the feature-interaction layer. We propose a distilled search space to cover the desired architectures with fewer parameters. We then develop a progressive search algorithm for efficient search on the space and well capture the order-priority property in sparse prediction tasks. Experiments on three real-world benchmark datasets show promising results of PROFIT in both accuracy and efficiency. Further studies validate the feasibility of our designed search space and search algorithm.
****
Local Explanation of Dialogue Response Generation
Yi-Lin Tuan, Connor Pryor, Wenhu Chen, Lise Getoor, William Yang 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.
****
Scalable Inference in SDEs by Direct Matching of the Fokker–Planck–Kolmogorov Equation
Arno Solin, Ella Tamir, Prakhar Verma
Simulation-based techniques such as variants of stochastic Runge–Kutta are the de facto approach for inference with stochastic differential equations (SDEs) in machine learning. These methods are general-purpose and used with parametric and non-parametric models, and neural SDEs. Stochastic Runge–Kutta relies on the use of sampling schemes that can be inefficient in high dimensions. We address this issue by revisiting the classical SDE literature and derive direct approximations to the (typically intractable) Fokker–Planck–Kolmogorov equation by matching moments. We show how this workflow is fast, scales to high-dimensional latent spaces, and is applicable to scarce-data applications, where a non-parametric SDE with a driving Gaussian process velocity field specifies the model.
****
The Complexity of Bayesian Network Learning: Revisiting the Superstructure
Robert Ganian, Viktoriia Korchemna
We investigate the parameterized complexity of Bayesian Network Structure Learning (BNSL), a classical problem that has received significant attention in empiri

cal but also purely theoretical studies. We follow up on previous works that hav
e analyzed the complexity of BNSL w.r.t. the so-called superstructure of the inp
ut. While known results imply that BNSL is unlikely to be fixed-parameter tracta
ble even when parameterized by the size of a vertex cover in the superstructure,
 here we show that a different kind of parameterization - notably by the size of
 a feedback edge set - yields fixed-parameter tractability. We proceed by showin
g that this result can be strengthened to a localized version of the feedback ed
ge set, and provide corresponding lower bounds that complement previous results
to provide a complexity classification of BNSL w.r.t. virtually all well-studied
 graph parameters.We then analyze how the complexity of BNSL depends on the repr
esentation of the input. In particular, while the bulk of past theoretical work
on the topic assumed the use of the so-called non-zero representation, here we p
rove that if an additive representation can be used instead then BNSL becomes fi
xed-parameter tractable even under significantly milder restrictions to the supe
rstructure, notably when parameterized by the treewidth alone. Last but not leas
t, we show how our results can be extended to the closely related problem of Pol
ytree Learning.
****
Fast Tucker Rank Reduction for Non-Negative Tensors Using Mean-Field Approximati
on
Kazu Ghalamkari, Mahito Sugiyama
We present an efficient low-rank approximation algorithm for non-negative tensor
s. The algorithm is derived from our two findings: First, we show that rank-1 a
pproximation for tensors can be viewed as a mean-field approximation by treating
 each tensor as a probability distribution. Second, we theoretically provide a s
ufficient condition for distribution parameters to reduce Tucker ranks of tensor
s; interestingly, this sufficient condition can be achieved by iterative applica
tion of the mean-field approximation. Since the mean-field approximation is alwa
ys given as a closed formula, our findings lead to a fast low-rank approximation
 algorithm without using a gradient method. We empirically demonstrate that our
algorithm is faster than the existing non-negative Tucker rank reduction methods
 and achieves competitive or better approximation of given tensors.
****
Learning Stochastic Majority Votes by Minimizing a PAC-Bayes Generalization Boun
d
Valentina Zantedeschi, Paul Viallard, Emilie Morvant, Rémi Emonet, Amaury Habrar
d, Pascal Germain, Benjamin Guedj
We investigate a stochastic counterpart of majority votes over finite ensembles
of classifiers, and study its generalization properties. While our approach hold
s for arbitrary distributions, we instantiate it with Dirichlet distributions: t
his allows for a closed-form and differentiable expression for the expected risk
, which then turns the generalization bound into a tractable training objective.
The resulting stochastic majority vote learning algorithm achieves state-of-the-
art accuracy and benefits from (non-vacuous) tight generalization bounds, in a s
eries of numerical experiments when compared to competing algorithms which also
minimize PAC-Bayes objectives -- both with uninformed (data-independent) and inf
ormed (data-dependent) priors.
****
Numerical influence of ReLU’(0) on backpropagation
David Bertoin, Jérôme Bolte, Sébastien Gerchinovitz, Edouard Pauwels
In theory, the choice of ReLU(0) in [0, 1] for a neural network has a negligible
 influence both on backpropagation and training. Yet, in the real world, 32 bits
 default precision combined with the size of deep learning problems makes it a h
yperparameter of training methods. We investigate the importance of the value of
 ReLU'(0) for several precision levels (16, 32, 64 bits), on various networks (f
ully connected, VGG, ResNet) and datasets (MNIST, CIFAR10, SVHN, ImageNet). We o
bserve considerable variations of backpropagation outputs which occur around hal
f of the time in 32 bits precision. The effect disappears with double precision,
 while it is systematic at 16 bits. For vanilla SGD training, the choice ReLU'(0
) = 0 seems to be the most efficient. For our experiments on ImageNet the gain i

n test accuracy over ReLU'(0) = 1 was more than 10 points (two runs). We also ev
idence that reconditioning approaches as batch-norm or ADAM tend to buffer the i
nfluence of ReLU'(0)’s value. Overall, the message we convey is that algorithmic
 differentiation of nonsmooth problems potentially hides parameters that could b
e tuned advantageously.
****
A Contrastive Learning Approach for Training Variational Autoencoder Priors
Jyoti Aneja, Alex Schwing, Jan Kautz, Arash Vahdat
Variational autoencoders (VAEs) are one of the powerful likelihood-based generat
ive models with applications in many domains. However, they struggle to generate
 high-quality images, especially when samples are obtained from the prior withou
t any tempering. One explanation for VAEs' poor generative quality is the prior
hole problem: the prior distribution fails to match the aggregate approximate po
sterior. Due to this mismatch, there exist areas in the latent space with high d
ensity under the prior that do not correspond to any encoded image. Samples from
 those areas are decoded to corrupted images. To tackle this issue, we propose a
n energy-based prior defined by the product of a base prior distribution and a r
eweighting factor, designed to bring the base closer to the aggregate posterior.
 We train the reweighting factor by noise contrastive estimation, and we general
ize it to hierarchical VAEs with many latent variable groups. Our experiments co
nfirm that the proposed noise contrastive priors improve the generative performa
nce of state-of-the-art VAEs by a large margin on the MNIST, CIFAR-10, CelebA 64
, and CelebA HQ 256 datasets. Our method is simple and can be applied to a wide
variety of VAEs to improve the expressivity of their prior distribution.
****
What training reveals about neural network complexity
Andreas Loukas, Marinos Poiitis, Stefanie Jegelka
This work explores the Benevolent Training Hypothesis (BTH) which argues that th
e complexity of the function a deep neural network (NN) is learning can be deduc
ed by its training dynamics. Our analysis provides evidence for BTH by relating
the NN's Lipschitz constant at different regions of the input space with the beh
avior of the stochastic training procedure. We first observe that the Lipschitz
 constant close to the training data affects various aspects of the parameter tr
ajectory, with more complex networks having a longer trajectory, bigger variance
, and often veering further from their initialization. We then show that NNs who
se 1st layer bias is trained more steadily (i.e., slowly and with little variati
on) have bounded complexity even in regions of the input space that are far from
 any training point. Finally, we find that steady training with Dropout implies
a training- and data-dependent generalization bound that grows poly-logarithmica
lly with the number of parameters. Overall, our results support the intuition th
at good training behavior can be a useful bias towards good generalization.
****
Class-agnostic Reconstruction of Dynamic Objects from Videos
Zhongzheng Ren, Xiaoming Zhao, Alex Schwing
We introduce REDO, a class-agnostic framework to REconstruct the Dynamic Objects
 from RGBD or calibrated videos. Compared to prior work, our problem setting is
more realistic yet more challenging for three reasons: 1) due to occlusion or ca
mera settings an object of interest may never be entirely visible, but we aim to
 reconstruct the complete shape; 2) we aim to handle different object dynamics i
ncluding rigid motion, non-rigid motion, and articulation; 3) we aim to reconstr
uct different categories of objects with one unified framework. To addre
ss these challenges, we develop two novel modules. First, we introduce a canon
ical 4D implicit function which is pixel-aligned with aggregated temporal visual
 cues. Second, we develop a 4D transformation module which captures object dynam
ics to support temporal propagation and aggregation. We study the efficacy of RE
DO in extensive experiments on synthetic RGBD video datasets SAIL-VOS 3D and Def
ormingThings4D++, and on real-world video data 3DPW. We find REDO outperforms s
tate-of-the-art dynamic reconstruction methods by a margin. In ablation studies
we validate each developed component.
****

Unique sparse decomposition of low rank matrices
Dian Jin, Xin Bing, Yuqian 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.
****
Neighborhood Reconstructing Autoencoders
Yonghyeon LEE, Hyeokjun Kwon, Frank Park
Vanilla autoencoders often produce manifolds that overfit to noisy training data, or have the wrong local connectivity and geometry. Autoencoder regularization techniques, e.g., the denoising autoencoder, have had some success in reducing overfitting, whereas recent graph-based methods that exploit local connectivity information provided by neighborhood graphs have had some success in mitigating local connectivity errors. Neither of these two approaches satisfactorily reduce both overfitting and connectivity errors; moreover, graph-based methods typically involve considerable preprocessing and tuning. To simultaneously address the two issues of overfitting and local connectivity, we propose a new graph-based autoencoder, the Neighborhood Reconstructing Autoencoder (NRAE). Unlike existing graph-based methods that attempt to encode the training data to some prescribed latent space distribution -- one consequence being that only the encoder is the object of the regularization -- NRAE merges local connectivity information contained in the neighborhood graphs with local quadratic approximations of the decoder function to formulate a new neighborhood reconstruction loss. Compared to existing graph-based methods, our new loss function is simple and easy to implement, and the resulting algorithm is scalable and computationally efficient; the only required preprocessing step is the construction of the neighborhood graph. Extensive experiments with standard datasets demonstrate that, compared to existing methods, NRAE improves both overfitting and local connectivity in the learned manifold, in some cases by significant margins. Code for NRAE is available at https://github.com/Gabe-YHLee/NRAE-public.
****
TopicNet: Semantic Graph-Guided Topic Discovery
Zhibin Duan, Yi.shi Xu, Bo Chen, dongsheng wang, Chaojie Wang, Mingyuan Zhou
Existing deep hierarchical topic models are able to extract semantically meaningful topics from a text corpus in an unsupervised manner and automatically organize them into a topic hierarchy. However, it is unclear how to incorporate prior belief such as knowledge graph to guide the learning of the topic hierarchy. To address this issue, we introduce TopicNet as a deep hierarchical topic model that can inject prior structural knowledge as inductive bias to influence the learning. TopicNet represents each topic as a Gaussian-distributed embedding vector, projects the topics of all layers into a shared embedding space, and explores both the symmetric and asymmetric similarities between Gaussian embedding vectors to incorporate prior semantic hierarchies. With a variational auto-encoding inference network, the model parameters are optimized by minimizing the evidence lower bound and supervised loss via stochastic gradient descent. Experiments on widely used benchmark show that TopicNet outperforms related deep topic models on discovering deeper interpretable topics and mining better document representations.
****
(Almost) Free Incentivized Exploration from Decentralized Learning Agents
Chengshuai Shi, Haifeng Xu, Wei Xiong, Cong Shen
Incentivized exploration in multi-armed bandits (MAB) has witnessed increasing interests and many progresses in recent years, where a principal offers bonuses to agents to do explorations on her behalf. However, almost all existing studies are confined to temporary myopic agents. In this work, we break this barrier and study incentivized exploration with multiple and long-term strategic agents, who have more complicated behaviors that often appear in real-world applications. An important observation of this work is that strategic agents' intrinsic needs of learning benefit (instead of harming) the principal's explorations by providi

ng "free pulls". Moreover, it turns out that increasing the population of agents significantly lowers the principal's burden of incentivizing. The key and somewhat surprising insight revealed from our results is that when there are sufficiently many learning agents involved, the exploration process of the principal can be (almost) free. Our main results are built upon three novel components which may be of independent interest: (1) a simple yet provably effective incentive-provision strategy; (2) a carefully crafted best arm identification algorithm for rewards aggregated under unequal confidences; (3) a high-probability finite-time lower bound of UCB algorithms. Experimental results are provided to complement the theoretical analysis.
****
Combining Recurrent, Convolutional, and Continuous-time Models with Linear State Space Layers
Albert Gu, Isys Johnson, Karan Goel, Khaled Saab, Tri Dao, Atri Rudra, Christopher Ré
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Revisiting Hilbert-Schmidt Information Bottleneck for Adversarial Robustness
Zifeng Wang, Tong Jian, Aria Masoomi, Stratis Ioannidis, Jennifer Dy
We investigate the HSIC (Hilbert-Schmidt independence criterion) bottleneck as a regularizer for learning an adversarially robust deep neural network classifier. In addition to the usual cross-entropy loss, we add regularization terms for every intermediate layer to ensure that the latent representations retain useful information for output prediction while reducing redundant information. We show that the HSIC bottleneck enhances robustness to adversarial attacks both theoretically and experimentally. In particular, we prove that the HSIC bottleneck regularizer reduces the sensitivity of the classifier to adversarial examples. Our experiments on multiple benchmark datasets and architectures demonstrate that incorporating an HSIC bottleneck regularizer attains competitive natural accuracy and improves adversarial robustness, both with and without adversarial examples during training. Our code and adversarially robust models are publicly available.
****
T-LoHo: A Bayesian Regularization Model for Structured Sparsity and Smoothness on Graphs
Changwoo Lee, Zhao Tang Luo, Huiyan Sang
Graphs have been commonly used to represent complex data structures. In models dealing with graph-structured data, multivariate parameters may not only exhibit sparse patterns but have structured sparsity and smoothness in the sense that both zero and non-zero parameters tend to cluster together. We propose a new prior for high-dimensional parameters with graphical relations, referred to as the Tree-based Low-rank Horseshoe (T-LoHo) model, that generalizes the popular univariate Bayesian horseshoe shrinkage prior to the multivariate setting to detect structured sparsity and smoothness simultaneously. The T-LoHo prior can be embedded in many high-dimensional hierarchical models. To illustrate its utility, we apply it to regularize a Bayesian high-dimensional regression problem where the regression coefficients are linked by a graph, so that the resulting clusters have flexible shapes and satisfy the cluster contiguity constraint with respect to the graph. We design an efficient Markov chain Monte Carlo algorithm that delivers full Bayesian inference with uncertainty measures for model parameters such as the number of clusters. We offer theoretical investigations of the clustering effects and posterior concentration results. Finally, we illustrate the performance of the model with simulation studies and a real data application for anomaly detection on a road network. The results indicate substantial improvements over other competing methods such as the sparse fused lasso.
****
The Utility of Explainable AI in Ad Hoc Human-Machine Teaming
Rohan Paleja, Muyleng Ghuy, Nadun Ranawaka Arachchige, Reed Jensen, Matthew Gomb

olay
Requests for name changes in the electronic proceedings will be accepted with 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.
****
Subgoal Search For Complex Reasoning Tasks
Konrad Czechowski, Tomasz Odrzygó■d■, Marek Zbysi■ski, Micha■ Zawalski, Krzyszto
f Olejnik, Yuhuai Wu, ■ukasz Kuci■ski, Piotr Mi■o■
Requests for name changes in the electronic proceedings will be accepted with 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.
****
MCMC Variational Inference via Uncorrected Hamiltonian Annealing
Tomas Geffner, Justin Domke
Given an unnormalized target distribution we want to obtain approximate samples
from it and a tight lower bound on its (log) normalization constant log Z. Annea
led Importance Sampling (AIS) with Hamiltonian MCMC is a powerful method that ca
n be used to do this. Its main drawback is that it uses non-differentiable trans
ition kernels, which makes tuning its many parameters hard. We propose a framewo
rk to use an AIS-like procedure with Uncorrected Hamiltonian MCMC, called Uncorr
ected Hamiltonian Annealing. Our method leads to tight and differentiable lower
bounds on log Z. We show empirically that our method yields better performances
than other competing approaches, and that the ability to tune its parameters usi
ng reparameterization gradients may lead to large performance improvements.
****
Landmark-RxR: Solving Vision-and-Language Navigation with Fine-Grained Alignment
 Supervision
Keji He, Yan Huang, Qi Wu, Jianhua Yang, Dong An, Shuanglin Sima, Liang Wang
In Vision-and-Language Navigation (VLN) task, an agent is asked to navigate insi
de 3D indoor environments following given instructions. Cross-modal alignment is
 one of the most critical challenges in VLN because the predicted trajectory nee
ds to match the given instruction accurately. In this paper, we address the cros
s-modal alignment challenge from the perspective of fine-grain. Firstly, to alle
viate weak cross-modal alignment supervision from coarse-grained data, we introd
uce a human-annotated fine-grained VLN dataset, namely Landmark-RxR. Secondly, t
o further enhance local cross-modal alignment under fine-grained supervision, we
 investigate the focal-oriented rewards with soft and hard forms, by focusing on
 the critical points sampled from fine-grained Landmark-RxR. Moreover, to fully
evaluate the navigation process, we also propose a re-initialization mechanism t
hat makes metrics insensitive to difficult points, which can cause the agent to
deviate from the correct trajectories. Experimental results show that our agent
has superior navigation performance on Landmark-RxR, en-RxR and R2R. Our dataset
 and code are available at https://github.com/hekj/Landmark-RxR.
****
A Winning Hand: Compressing Deep Networks Can Improve Out-of-Distribution Robust
ness
James Diffenderfer, Brian Bartoldson, Shreya Chaganti, Jize Zhang, Bhavya Kailkh
ura
Successful adoption of deep learning (DL) in the wild requires models to be: (1)
 compact, (2) accurate, and (3) robust to distributional shifts. Unfortunately,
efforts towards simultaneously meeting these requirements have mostly been unsuc
cessful. This raises an important question: Is the inability to create Compact,
Accurate, and Robust Deep neural networks (CARDs) fundamental? To answer this qu
estion, we perform a large-scale analysis of popular model compression technique
s which uncovers several intriguing patterns. Notably, in contrast to traditiona
l pruning approaches (e.g., fine tuning and gradual magnitude pruning), we find
that ``lottery ticket-style'' approaches can surprisingly be used to produce CAR
Ds, including binary-weight CARDs. Specifically, we are able to create extremely

 compact CARDs that, compared to their larger counterparts, have similar test ac
curacy and matching (or better) robustness---simply by pruning and (optionally)
quantizing. Leveraging the compactness of CARDs, we develop a simple domain-adap
tive test-time ensembling approach (CARD-Decks) that uses a gating module to dyn
amically select appropriate CARDs from the CARD-Deck based on their spectral-sim
ilarity with test samples. The proposed approach builds a "winning hand'' of CAR
Ds that establishes a new state-of-the-art (on RobustBench) on CIFAR-10-C accura
cies (i.e., 96.8% standard and 92.75% robust) and CIFAR-100-C accuracies (80.6%
standard and 71.3% robust) with better memory usage than non-compressed baseline
s (pretrained CARDs and CARD-Decks available at https://github.com/RobustBench/r
obustbench). Finally, we provide theoretical support for our empirical findings.

****
On the Importance of Gradients for Detecting Distributional Shifts in the Wild
Rui Huang, Andrew Geng, Yixuan Li
Detecting out-of-distribution (OOD) data has become a critical component in ensu
ring the safe deployment of machine learning models in the real world. Existing
OOD detection approaches primarily rely on the output or feature space for deriv
ing OOD scores, while largely overlooking information from the gradient space. I
n this paper, we present GradNorm, a simple and effective approach for detecting
 OOD inputs by utilizing information extracted from the gradient space. GradNorm
 directly employs the vector norm of gradients, backpropagated from the KL diver
gence between the softmax output and a uniform probability distribution. Our key
 idea is that the magnitude of gradients is higher for in-distribution (ID) data
 than that for OOD data, making it informative for OOD detection. GradNorm demon
strates superior performance, reducing the average FPR95 by up to 16.33% compare
d to the previous best method.
****
Iterative Methods for Private Synthetic Data: Unifying Framework and New Methods
Terrance Liu, Giuseppe Vietri, Steven Z. Wu
We study private synthetic data generation for query release, where the goal is
to construct a sanitized version of a sensitive dataset, subject to differential
 privacy, that approximately preserves the answers to a large collection of stat
istical queries. We first present an algorithmic framework that unifies a long l
ine of iterative algorithms in the literature. Under this framework, we propose
two new methods. The first method, private entropy projection (PEP), can be view
ed as an advanced variant of MWEM that adaptively reuses past query measurements
 to boost accuracy. Our second method, generative networks with the exponential
mechanism (GEM), circumvents computational bottlenecks in algorithms such as MWE
M and PEP by optimizing over generative models parameterized by neural networks,
 which capture a rich family of distributions while enabling fast gradient-based
 optimization. We demonstrate that PEP and GEM empirically outperform existing a
lgorithms. Furthermore, we show that GEM nicely incorporates prior information f
rom public data while overcoming limitations of PMW^Pub, the existing state-of-t
he-art method that also leverages public data.
****
Understanding End-to-End Model-Based Reinforcement Learning Methods as Implicit
Parameterization
Clement Gehring, Kenji Kawaguchi, Jiaoyang Huang, Leslie Kaelbling
Estimating the per-state expected cumulative rewards is a critical aspect of rei
nforcement learning approaches, however the experience is obtained, but standard
 deep neural-network function-approximation methods are often inefficient in thi
s setting. An alternative approach, exemplified by value iteration networks, is
to learn transition and reward models of a latent Markov decision process whose
value predictions fit the data. This approach has been shown empirically to conv
erge faster to a more robust solution in many cases, but there has been little t
heoretical study of this phenomenon. In this paper, we explore such implicit rep
resentations of value functions via theory and focused experimentation. We prove
 that, for a linear parametrization, gradient descent converges to global optima
 despite non-linearity and non-convexity introduced by the implicit representati

on. Furthermore, we derive convergence rates for both cases which allow us to identify conditions under which stochastic gradient descent (SGD) with this implicit representation converges substantially faster than its explicit counterpart. Finally, we provide empirical results in some simple domains that illustrate the theoretical findings.
****
Mirror Langevin Monte Carlo: the Case Under Isoperimetry
Qijia Jiang
Motivated by the connection between sampling and optimization, we study a mirror descent analogue of Langevin dynamics and analyze three different discretization schemes, giving nonasymptotic convergence rate under functional inequalities such as Log-Sobolev in the corresponding metric. Compared to the Euclidean setting, the result reveals intricate relationship between the underlying geometry and the target distribution and suggests that care might need to be taken in order for the discretized algorithm to achieve vanishing bias with diminishing stepsize for sampling from potentials under weaker smoothness/convexity regularity conditions.
****
Do Different Tracking Tasks Require Different Appearance Models?
Zhongdao Wang, Hengshuang Zhao, Ya-Li Li, Shengjin Wang, Philip Torr, Luca Bertinetto
Tracking objects of interest in a video is one of the most popular and widely applicable problems in computer vision. However, with the years, a Cambrian explosion of use cases and benchmarks has fragmented the problem in a multitude of different experimental setups. As a consequence, the literature has fragmented too, and now novel approaches proposed by the community are usually specialised to fit only one specific setup. To understand to what extent this specialisation is necessary, in this work we present UniTrack, a solution to address five different tasks within the same framework. UniTrack consists of a single and task-agnostic appearance model, which can be learned in a supervised or self-supervised fashion, and multiple ``heads'' that address individual tasks and do not require training. We show how most tracking tasks can be solved within this framework, and that the same appearance model can be successfully used to obtain results that are competitive against specialised methods for most of the tasks considered. The framework also allows us to analyse appearance models obtained with the most recent self-supervised methods, thus extending their evaluation and comparison to a larger variety of important problems.
****
Towards robust vision by multi-task learning on monkey visual cortex
Shahd Safarani, Arne Nix, Konstantin Willeke, Santiago Cadena, Kelli Restivo, George Denfield, Andreas Tolias, Fabian Sinz
Deep neural networks set the state-of-the-art across many tasks in computer vision, but their generalization ability to simple image distortions is surprisingly fragile. In contrast, the mammalian visual system is robust to a wide range of perturbations. Recent work suggests that this generalization ability can be explained by useful inductive biases encoded in the representations of visual stimuli throughout the visual cortex. Here, we successfully leveraged these inductive biases with a multi-task learning approach: we jointly trained a deep network to perform image classification and to predict neural activity in macaque primary visual cortex (V1) in response to the same natural stimuli. We measured the out-of-distribution generalization abilities of our resulting network by testing its robustness to common image distortions. We found that co-training on monkey V1 data indeed leads to increased robustness despite the absence of those distortions during training. Additionally, we showed that our network's robustness is often very close to that of an Oracle network where parts of the architecture are directly trained on noisy images. Our results also demonstrated that the network's representations become more brain-like as their robustness improves. Using a novel constrained reconstruction analysis, we investigated what makes our brain-regularized network more robust. We found that our monkey co-trained network is more sensitive to content than noise when compared to a Baseline network that we

trained for image classification alone. Using DeepGaze-predicted saliency maps f
or ImageNet images, we found that the monkey co-trained network tends to be more
 sensitive to salient regions in a scene, reminiscent of existing theories on th
e role of V1 in the detection of object borders and bottom-up saliency. Overall,
 our work expands the promising research avenue of transferring inductive biases
 from biological to artificial neural networks on the representational level, an
d provides a novel analysis of the effects of our transfer.
****
Arbitrary Conditional Distributions with Energy
Ryan Strauss, Junier B. Oliva
Requests for name changes in the electronic proceedings will be accepted with 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 Domain Invariant Representations in Goal-conditioned Block MDPs
Beining Han, Chongyi Zheng, Harris Chan, Keiran Paster, Michael Zhang, Jimmy Ba
Deep Reinforcement Learning (RL) is successful in solving many complex Markov De
cision Processes (MDPs) problems. However, agents often face unanticipated envir
onmental changes after deployment in the real world. These changes are often spu
rious and unrelated to the underlying problem, such as background shifts for vis
ual input agents. Unfortunately, deep RL policies are usually sensitive to these
 changes and fail to act robustly against them. This resembles the problem of do
main generalization in supervised learning. In this work, we study this problem
for goal-conditioned RL agents. We propose a theoretical framework in the Block
MDP setting that characterizes the generalizability of goal-conditioned policies
 to new environments. Under this framework, we develop a practical method PA-Ske
wFit that enhances domain generalization. The empirical evaluation shows that ou
r goal-conditioned RL agent can perform well in various unseen test environments
, improving by 50\% over baselines.
****
Near-Optimal Multi-Perturbation Experimental Design for Causal Structure Learnin
g
Scott Sussex, Caroline Uhler, Andreas Krause
Causal structure learning is a key problem in many domains. Causal structures ca
n be learnt by performing experiments on the system of interest. We address the
largely unexplored problem of designing a batch of experiments that each simulta
neously intervene on multiple variables. While potentially more informative than
 the commonly considered single-variable interventions, selecting such intervent
ions is algorithmically much more challenging, due to the doubly-exponential com
binatorial search space over sets of composite interventions. In this paper, we
develop efficient algorithms for optimizing different objective functions quanti
fying the informativeness of a budget-constrained batch of experiments. By estab
lishing novel submodularity properties of these objectives, we provide approxima
tion guarantees for our algorithms. Our algorithms empirically perform superior
to both random interventions and algorithms that only select single-variable int
erventions.
****
Fuzzy Clustering with Similarity Queries
Wasim Huleihel, Arya Mazumdar, Soumyabrata Pal
Requests for name changes in the electronic proceedings will be accepted with no
 questions asked. However name changes may cause bibliographic tracking issues.
  Authors are asked to consider this carefully and discuss it with their co-auth
ors prior to requesting a name change in the electronic proceedings.
****
Improving black-box optimization in VAE latent space using decoder uncertainty
Pascal Notin, José Miguel Hernández-Lobato, Yarin Gal
Optimization in the latent space of variational autoencoders is a promising appr
oach to generate high-dimensional discrete objects that maximize an expensive bl
ack-box property (e.g., drug-likeness in molecular generation, function approxim

ation with arithmetic expressions). However, existing methods lack robustness as they may decide to explore areas of the latent space for which no data was available during training and where the decoder can be unreliable, leading to the generation of unrealistic or invalid objects. We propose to leverage the epistemic uncertainty of the decoder to guide the optimization process. This is not trivial though, as a naive estimation of uncertainty in the high-dimensional and structured settings we consider would result in high estimator variance. To solve this problem, we introduce an importance sampling-based estimator that provides more robust estimates of epistemic uncertainty. Our uncertainty-guided optimization approach does not require modifications of the model architecture nor the training process. It produces samples with a better trade-off between black-box objective and validity of the generated samples, sometimes improving both simultaneously. We illustrate these advantages across several experimental settings in digit generation, arithmetic expression approximation and molecule generation for drug design.
****

Sample Selection for Fair and Robust Training

Yuji Roh, Kangwook Lee, Steven Whang, Changho Suh

Fairness and robustness are critical elements of Trustworthy AI that need to be addressed together. Fairness is about learning an unbiased model while robustness is about learning from corrupted data, and it is known that addressing only one of them may have an adverse affect on the other. In this work, we propose a sample selection-based algorithm for fair and robust training. To this end, we formulate a combinatorial optimization problem for the unbiased selection of samples in the presence of data corruption. Observing that solving this optimization problem is strongly NP-hard, we propose a greedy algorithm that is efficient and effective in practice. Experiments show that our method obtains fairness and robustness that are better than or comparable to the state-of-the-art technique, both on synthetic and benchmark real datasets. Moreover, unlike other fair and robust training baselines, our algorithm can be used by only modifying the sampling step in batch selection without changing the training algorithm or leveraging additional clean data.
****

NeurWIN: Neural Whittle Index Network For Restless Bandits Via Deep RL

Khaled Nakhleh, Santosh Ganji, Ping-Chun Hsieh, I-Hong Hou, Srinivas Shakkottai

Whittle index policy is a powerful tool to obtain asymptotically optimal solutions for the notoriously intractable problem of restless bandits. However, finding the Whittle indices remains a difficult problem for many practical restless bandits with convoluted transition kernels. This paper proposes NeurWIN, a neural Whittle index network that seeks to learn the Whittle indices for any restless bandits by leveraging mathematical properties of the Whittle indices. We show that a neural network that produces the Whittle index is also one that produces the optimal control for a set of Markov decision problems. This property motivates using deep reinforcement learning for the training of NeurWIN. We demonstrate the utility of NeurWIN by evaluating its performance for three recently studied restless bandit problems.Our experiment results show that the performance of NeurWIN is significantly better than other RL algorithms.
****

Sageflow: Robust Federated Learning against Both Stragglers and Adversaries

Jungwuk Park, Dong-Jun Han, Minseok Choi, Jaekyun Moon

While federated learning (FL) allows efficient model training with local data at edge devices, among major issues still to be resolved are: slow devices known as stragglers and malicious attacks launched by adversaries.   While the presence of both of these issues raises serious concerns in practical FL systems, no known schemes or combinations of schemes effectively address them at the same time. We propose Sageflow, staleness-aware grouping with entropy-based filtering and loss-weighted averaging, to handle both stragglers and adversaries simultaneously. Model grouping and weighting according to staleness (arrival delay) provides robustness against stragglers, while entropy-based filtering and loss-weighted averaging, working in a highly complementary fashion at each grouping stage, cou

nter a wide range of adversary attacks. A theoretical bound is established to provide key insights into the convergence behavior of Sageflow. Extensive experimental results show that Sageflow outperforms various existing methods aiming to handle stragglers/adversaries.
****
Alias-Free Generative Adversarial Networks
Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, Timo Aila
We observe that despite their hierarchical convolutional nature, the synthesis process of typical generative adversarial networks depends on absolute pixel coordinates in an unhealthy manner. This manifests itself as, e.g., detail appearing to be glued to image coordinates instead of the surfaces of depicted objects. We trace the root cause to careless signal processing that causes aliasing in the generator network. Interpreting all signals in the network as continuous, we derive generally applicable, small architectural changes that guarantee that unwanted information cannot leak into the hierarchical synthesis process. The resulting networks match the FID of StyleGAN2 but differ dramatically in their internal representations, and they are fully equivariant to translation and rotation even at subpixel scales. Our results pave the way for generative models better suited for video and animation.
****
Noise2Score: Tweedie’s Approach to Self-Supervised Image Denoising without Clean Images
Kwanyoung Kim, Jong Chul Ye
Recently, there has been extensive research interest in training deep networks to denoise images without clean reference.However, the representative approaches such as Noise2Noise, Noise2Void, Stein's unbiased risk estimator (SURE), etc. seem to differ from one another and it is difficult to find the coherent mathematical structure. To address this, here we present a novel approach, called Noise2Score, which reveals a missing link in order to unite these seemingly different approaches.Specifically, we show that image denoising problems without clean images can be addressed by finding the mode of the posterior distribution and that the Tweedie's formula offers an explicit solution through the score function (i.e. the gradient of loglikelihood). Our method then uses the recent finding that the score function can be stably estimated from the noisy images using the amortized residual denoising autoencoder, the method of which is closely related to Noise2Noise or Nose2Void. Our Noise2Score approach is so universal that the same network training can be used to remove noises from images that are corrupted by any exponential family distributions and noise parameters. Using extensive experiments with Gaussian, Poisson, and Gamma noises, we show that Noise2Score significantly outperforms the state-of-the-art self-supervised denoising methods in the benchmark data set such as (C)BSD68, Set12, and Kodak, etc.
****
Continuous Mean-Covariance Bandits
Yihan Du, Siwei Wang, Zhixuan Fang, Longbo Huang
Existing risk-aware multi-armed bandit models typically focus on risk measures of individual options such as variance. As a result, they cannot be directly applied to important real-world online decision making problems with correlated options. In this paper, we propose a novel Continuous Mean-Covariance Bandit (CMCB) model to explicitly take into account option correlation. Specifically, in CMCB, there is a learner who sequentially chooses weight vectors on given options and observes random feedback according to the decisions. The agent's objective is to achieve the best trade-off between reward and risk, measured with option covariance. To capture different reward observation scenarios in practice, we consider three feedback settings, i.e., full-information, semi-bandit and full-bandit feedback. We propose novel algorithms with optimal regrets (within logarithmic factors), and provide matching lower bounds to validate their optimalities. The experimental results also demonstrate the superiority of our algorithms. To the best of our knowledge, this is the first work that considers option correlation in risk-aware bandits and explicitly quantifies how arbitrary covariance struct

ures impact the learning performance.The novel analytical techniques we developed for exploiting the estimated covariance to build concentration and bounding the risk of selected actions based on sampling strategy properties can likely find applications in other bandit analysis and be of independent interests.
****

Dynamic Visual Reasoning by Learning Differentiable Physics Models from Video and Language
Mingyu Ding, Zhenfang Chen, Tao Du, Ping Luo, Josh Tenenbaum, Chuang Gan
In this work, we propose a unified framework, called Visual Reasoning with Differ-entiable Physics (VRDP), that can jointly learn visual concepts and infer physics models of objects and their interactions from videos and language. This is achieved by seamlessly integrating three components: a visual perception module, a concept learner, and a differentiable physics engine. The visual perception module parses each video frame into object-centric trajectories and represents them as latent scene representations. The concept learner grounds visual concepts (e.g., color, shape, and material) from these object-centric representations based on the language, thus providing prior knowledge for the physics engine. The differentiable physics model, implemented as an impulse-based differentiable rigid-body simulator, performs differentiable physical simulation based on the grounded concepts to infer physical properties, such as mass, restitution, and velocity, by fitting the simulated trajectories into the video observations. Consequently, these learned concepts and physical models can explain what we have seen and imagine what is about to happen in future and counterfactual scenarios. Integrating differentiable physics into the dynamic reasoning framework offers several appealing benefits. More accurate dynamics prediction in learned physics models enables state-of-the-art performance on both synthetic and real-world benchmarks while still maintaining high transparency and interpretability; most notably, VRDP improves the accuracy of predictive and counterfactual questions by 4.5% and 11.5% compared to its best counterpart. VRDP is also highly data-efficient: physical parameters can be optimized from very few videos, and even a single video can be sufficient. Finally, with all physical parameters inferred, VRDP can quickly learn new concepts from a few examples.
****

Solving Soft Clustering Ensemble via k-Sparse Discrete Wasserstein Barycenter
Ruizhe Qin, Mengying Li, Hu Ding
Clustering ensemble is one of the most important problems in ensemble learning. Though it has been extensively studied in the past decades, the existing methods often suffer from the issues like high computational complexity and the difficulty on understanding the consensus. In this paper, we study the more general soft clustering ensemble problem where each individual solution is a soft clustering. We connect it to the well-known discrete Wasserstein barycenter problem in geometry. Based on some novel geometric insights in high dimensions, we propose the sampling-based algorithms with provable quality guarantees. We also provide the systematical analysis on the consensus of our model. Finally, we conduct the experiments to evaluate our proposed algorithms.
****

Bayesian Adaptation for Covariate Shift
Aurick Zhou, Sergey Levine
When faced with distribution shift at test time, deep neural networks often make inaccurate predictions with unreliable uncertainty estimates.While improving the robustness of neural networks is one promising approach to mitigate this issue, an appealing alternate to robustifying networks against all possible test-time shifts is to instead directly adapt them to unlabeled inputs from the particular distribution shift we encounter at test time.However, this poses a challenging question: in the standard Bayesian model for supervised learning, unlabeled inputs are conditionally independent of model parameters when the labels are unobserved, so what can unlabeled data tell us about the model parameters at test-time? In this paper, we derive a Bayesian model that provides for a well-defined relationship between unlabeled inputs under distributional shift and model parameters, and show how approximate inference in this model can be instantiated with a

simple regularized entropy minimization procedure at test-time. We evaluate our method on a variety of distribution shifts for image classification, including image corruptions, natural distribution shifts, and domain adaptation settings, and show that our method improves both accuracy and uncertainty estimation.
****
Perturb-and-max-product: Sampling and learning in discrete energy-based models
Miguel Lazaro-Gredilla, Antoine Dedieu, Dileep George
Perturb-and-MAP offers an elegant approach to approximately sample from a energy-based model (EBM) by computing the maximum-a-posteriori (MAP) configuration of a perturbed version of the model. Sampling in turn enables learning. However, this line of research has been hindered by the general intractability of the MAP computation. Very few works venture outside tractable models, and when they do, they use linear programming approaches, which as we will show, have several limitations. In this work we present perturb-and-max-product (PMP), a parallel and scalable mechanism for sampling and learning in discrete EBMs. Models can be arbitrary as long as they are built using tractable factors. We show that (a) for Ising models, PMP is orders of magnitude faster than Gibbs and Gibbs-with-Gradients (GWG) at learning and generating samples of similar or better quality; (b) PMP is able to learn and sample from RBMs; (c) in a large, entangled graphical model in which Gibbs and GWG fail to mix, PMP succeeds.
****
Towards Unifying Behavioral and Response Diversity for Open-ended Learning in Zero-sum Games
Xiangyu Liu, Hangtian Jia, Ying Wen, Yujing Hu, Yingfeng Chen, Changjie Fan, ZHIPENG HU, Yaodong Yang
Measuring and promoting policy diversity is critical for solving games with strong non-transitive dynamics where strategic cycles exist, and there is no consistent winner (e.g., Rock-Paper-Scissors). With that in mind, maintaining a pool of diverse policies via open-ended learning is an attractive solution, which can generate auto-curricula to avoid being exploited. However, in conventional open-ended learning algorithms, there are no widely accepted definitions for diversity, making it hard to construct and evaluate the diverse policies. In this work, we summarize previous concepts of diversity and work towards offering a unified measure of diversity in multi-agent open-ended learning to include all elements in Markov games, based on both Behavioral Diversity (BD) and Response Diversity (RD). At the trajectory distribution level, we re-define BD in the state-action space as the discrepancies of occupancy measures. For the reward dynamics, we propose RD to characterize diversity through the responses of policies when encountering different opponents. We also show that many current diversity measures fall in one of the categories of BD or RD but not both. With this unified diversity measure, we design the corresponding diversity-promoting objective and population effectivity when seeking the best responses in open-ended learning. We validate our methods in both relatively simple games like matrix game, non-transitive mixture model, and the complex \textit{Google Research Football} environment. The population found by our methods reveals the lowest exploitability, highest population effectivity in matrix game and non-transitive mixture model, as well as the largest goal difference when interacting with opponents of various levels in \textit{Google Research Football}.
****
Towards Better Understanding of Training Certifiably Robust Models against Adversarial Examples
Sungyoon Lee, Woojin Lee, Jinseong Park, Jaewook Lee
We study the problem of training certifiably robust models against adversarial examples. Certifiable training minimizes an upper bound on the worst-case loss over the allowed perturbation, and thus the tightness of the upper bound is an important factor in building certifiably robust models. However, many studies have shown that Interval Bound Propagation (IBP) training uses much looser bounds but outperforms other models that use tighter bounds. We identify another key factor that influences the performance of certifiable training: \textit{smoothness of the loss landscape}. We find significant differences in the loss landscapes acr

oss many linear relaxation-based methods, and that the current state-of-the-arts method often has a landscape with favorable optimization properties. Moreover, to test the claim, we design a new certifiable training method with the desired properties. With the tightness and the smoothness, the proposed method achieves a decent performance under a wide range of perturbations, while others with only one of the two factors can perform well only for a specific range of perturbations. Our code is available at \url{https://github.com/sungyoon-lee/LossLandscapeMatters}.
****
Mitigating Covariate Shift in Imitation Learning via Offline Data With Partial Coverage
Jonathan Chang, Masatoshi Uehara, Dhruv Sreenivas, Rahul Kidambi, Wen Sun
This paper studies offline Imitation Learning (IL) where an agent learns to imitate an expert demonstrator without additional online environment interactions. Instead, the learner is presented with a static offline dataset of state-action-next state triples from a potentially less proficient behavior policy. We introduce Model-based IL from Offline data (MILO): an algorithmic framework that utilizes the static dataset to solve the offline IL problem efficiently both in theory and in practice. In theory, even if the behavior policy is highly sub-optimal compared to the expert, we show that as long as the data from the behavior policy provides sufficient coverage on the expert state-action traces (and with no necessity for a global coverage over the entire state-action space), MILO can provably combat the covariate shift issue in IL. Complementing our theory results, we also demonstrate that a practical implementation of our approach mitigates covariate shift on benchmark MuJoCo continuous control tasks. We demonstrate that with behavior policies whose performances are less than half of that of the expert, MILO still successfully imitates with an extremely low number of expert state-action pairs while traditional offline IL methods such as behavior cloning (BC) fail completely. Source code is provided at https://github.com/jdchang1/milo.
****
Global Filter Networks for Image Classification
Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, Jie Zhou
Recent advances in self-attention and pure multi-layer perceptrons (MLP) models for vision have shown great potential in achieving promising performance with fewer inductive biases. These models are generally based on learning interaction among spatial locations from raw data. The complexity of self-attention and MLP grows quadratically as the image size increases, which makes these models hard to scale up when high-resolution features are required. In this paper, we present the Global Filter Network (GFNet), a conceptually simple yet computationally efficient architecture, that learns long-term spatial dependencies in the frequency domain with log-linear complexity. Our architecture replaces the self-attention layer in vision transformers with three key operations: a 2D discrete Fourier transform, an element-wise multiplication between frequency-domain features and learnable global filters, and a 2D inverse Fourier transform. We exhibit favorable accuracy/complexity trade-offs of our models on both ImageNet and downstream tasks. Our results demonstrate that GFNet can be a very competitive alternative to transformer-style models and CNNs in efficiency, generalization ability and robustness. Code is available at https://github.com/raoyongming/GFNet
****
CAFE: Catastrophic Data Leakage in Vertical Federated Learning
Xiao Jin, Pin-Yu Chen, Chia-Yi Hsu, Chia-Mu Yu, Tianyi Chen
Recent studies show that private training data can be leaked through the gradients sharing mechanism deployed in distributed machine learning systems, such as federated learning (FL). Increasing batch size to complicate data recovery is often viewed as a promising defense strategy against data leakage. In this paper, we revisit this defense premise and propose an advanced data leakage attack with theoretical justification to efficiently recover batch data from the shared aggregated gradients. We name our proposed method as catastrophic data leakage in vertical federated learning (CAFE). Comparing to existing data leakage attacks, our extensive experimental results on vertical FL settings demonstrate the effecti

veness of CAFE to perform large-batch data leakage attack with improved data rec overy quality. We also propose a practical countermeasure to mitigate CAFE. Our results suggest that private data participated in standard FL, especially the ve rtical case, have a high risk of being leaked from the training gradients. Our a nalysis implies unprecedented and practical data leakage risks in those learning settings. The code of our work is available at https://github.com/DeRafael/CAFE .
****
Fault-Tolerant Federated Reinforcement Learning with Theoretical Guarantee
Xiaofeng Fan, Yining Ma, Zhongxiang Dai, Wei Jing, Cheston Tan, Bryan Kian Hsian g Low
The growing literature of Federated Learning (FL) has recently inspired Federate d Reinforcement Learning (FRL) to encourage multiple agents to federatively buil d a better decision-making policy without sharing raw trajectories. Despite its promising applications, existing works on FRL fail to I) provide theoretical ana lysis on its convergence, and II) account for random system failures and adversa rial attacks. Towards this end, we propose the first FRL framework the convergen ce of which is guaranteed and tolerant to less than half of the participating ag ents being random system failures or adversarial attackers. We prove that the sa mple efficiency of the proposed framework is guaranteed to improve with the numb er of agents and is able to account for such potential failures or attacks. All theoretical results are empirically verified on various RL benchmark tasks.
****
Compacter: Efficient Low-Rank Hypercomplex Adapter Layers
Rabeeh Karimi Mahabadi, James Henderson, Sebastian Ruder
Adapting large-scale pretrained language models to downstream tasks via fine-tun ing is the standard method for achieving state-of-the-art performance on NLP ben chmarks. However, fine-tuning all weights of models with millions or billions of parameters is sample-inefficient, unstable in low-resource settings, and wastef ul as it requires storing a separate copy of the model for each task. Recent wor k has developed parameter-efficient fine-tuning methods, but these approaches e ither still require a relatively large number of parameters or underperform stan dard fine-tuning. In this work, we propose Compacter, a method for fine-tuning l arge-scale language models with a better trade-off between task performance and the number of trainable parameters than prior work. Compacter accomplishes this by building on top of ideas from adapters, low-rank optimization, and parameteri zed hypercomplex multiplication layers.Specifically, Compacter inserts task-spec ific weight matrices into a pretrained model's weights, which are computed effic iently as a sum of Kronecker products between shared slow'' weights andfast'' ra nk-one matrices defined per Compacter layer. By only training 0.047% of a pretra ined model's parameters, Compacter performs on par with standard fine-tuning on GLUE and outperforms standard fine-tuning on SuperGLUE and low-resource settings . Our code is publicly available at https://github.com/rabeehk/compacter.
****
Distilling Image Classifiers in Object Detectors
Shuxuan Guo, Jose M. Alvarez, Mathieu Salzmann
Knowledge distillation constitutes a simple yet effective way to improve the per formance of a compact student network by exploiting the knowledge of a more powe rful teacher. Nevertheless, the knowledge distillation literature remains limite d to the scenario where the student and the teacher tackle the same task. Here, we investigate the problem of transferring knowledge not only across architectur es but also across tasks. To this end, we study the case of object detection and , instead of following the standard detector-to-detector distillation approach, introduce a classifier-to-detector knowledge transfer framework. In particular, we propose strategies to exploit the classification teacher to improve both the detector's recognition accuracy and localization performance. Our experiments on several detectors with different backbones demonstrate the effectiveness of our approach, allowing us to outperform the state-of-the-art detector-to-detector d istillation methods.
****

Subgroup Generalization and Fairness of Graph Neural Networks
Jiaqi Ma, Junwei Deng, Qiaozhu Mei
Despite enormous successful applications of graph neural networks (GNNs), theore tical understanding of their generalization ability, especially for node-level t asks where data are not independent and identically-distributed (IID), has been sparse. The theoretical investigation of the generalization performance is benef icial for understanding fundamental issues (such as fairness) of GNN models and designing better learning methods. In this paper, we present a novel PAC-Bayesia n analysis for GNNs under a non-IID semi-supervised learning setup. Moreover, we analyze the generalization performances on different subgroups of unlabeled nod es, which allows us to further study an accuracy-(dis)parity-style (un)fairness of GNNs from a theoretical perspective. Under reasonable assumptions, we demonst rate that the distance between a test subgroup and the training set can be a key factor affecting the GNN performance on that subgroup, which calls special atte ntion to the training node selection for fair learning. Experiments across multi ple GNN models and datasets support our theoretical results.
****
Scaling Neural Tangent Kernels via Sketching and Random Features
Amir Zandieh, Insu Han, Haim Avron, Neta Shoham, Chaewon Kim, Jinwoo Shin
The Neural Tangent Kernel (NTK) characterizes the behavior of infinitely-wide ne ural networks trained under least squares loss by gradient descent. Recent works also report that NTK regression can outperform finitely-wide neural networks tr ained on small-scale datasets. However, the computational complexity of kernel m ethods has limited its use in large-scale learning tasks. To accelerate learning with NTK, we design a near input-sparsity time approximation algorithm for NTK, by sketching the polynomial expansions of arc-cosine kernels: our sketch for th e convolutional counterpart of NTK (CNTK) can transform any image using a linear runtime in the number of pixels. Furthermore, we prove a spectral approximation guarantee for the NTK matrix, by combining random features (based on leverage s core sampling) of the arc-cosine kernels with a sketching algorithm. We benchmar k our methods on various large-scale regression and classification tasks and sho w that a linear regressor trained on our CNTK features matches the accuracy of e xact CNTK on CIFAR-10 dataset while achieving 150x speedup.
****
BatchQuant: Quantized-for-all Architecture Search with Robust Quantizer
Haoping Bai, Meng Cao, Ping Huang, Jiulong Shan
Requests for name changes in the electronic proceedings will be accepted with 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.
****
Long Short-Term Transformer for Online Action Detection
Mingze Xu, Yuanjun Xiong, Hao Chen, Xinyu Li, Wei Xia, Zhuowen Tu, Stefano Soatt o
We present Long Short-term TRansformer (LSTR), a temporal modeling algorithm for online action detection, which employs a long- and short-term memory mechanism to model prolonged sequence data. It consists of an LSTR encoder that dynamicall y leverages coarse-scale historical information from an extended temporal window (e.g., 2048 frames spanning of up to 8 minutes), together with an LSTR decoder that focuses on a short time window (e.g., 32 frames spanning 8 seconds) to mode l the fine-scale characteristics of the data. Compared to prior work, LSTR provi des an effective and efficient method to model long videos with fewer heuristics , which is validated by extensive empirical analysis. LSTR achieves state-of-the -art performance on three standard online action detection benchmarks, THUMOS'14 , TVSeries, and HACS Segment. Code has been made available at: https://xumingze0 308.github.io/projects/lstr.
****
Near Optimal Policy Optimization via REPS
Aldo Pacchiano, Jonathan N Lee, Peter Bartlett, Ofir Nachum
Since its introduction a decade ago, relative entropy policy search (REPS) has d

emonstrated successful policy learning on a number of simulated and real-world robotic domains, not to mention providing algorithmic components used by many recently proposed reinforcement learning (RL) algorithms. While REPS is commonly known in the community, there exist no guarantees on its performance when using stochastic and gradient-based solvers. In this paper we aim to fill this gap by providing guarantees and convergence rates for the sub-optimality of a policy learned using first-order optimization methods applied to the REPS objective. We first consider the setting in which we are given access to exact gradients and demonstrate how near-optimality of the objective translates to near-optimality of the policy. We then consider the practical setting of stochastic gradients, and introduce a technique that uses generative access to the underlying Markov decision process to compute parameter updates that maintain favorable convergence to the optimal regularized policy.
****

Self-Consistent Models and Values
Greg Farquhar, Kate Baumli, Zita Marinho, Angelos Filos, Matteo Hessel, Hado P. van Hasselt, David Silver
Learned models of the environment provide reinforcement learning (RL) agents with flexible ways of making predictions about the environment.Models enable planning, i.e. using more computation to improve value functions or policies, without requiring additional environment interactions.In this work, we investigate a way of augmenting model-based RL, by additionally encouraging a learned model and value function to be jointly \emph{self-consistent}.This lies in contrast to classic planning methods like Dyna, which only update the value function to be consistent with the model.We propose a number of possible self-consistency updates, study them empirically in both the tabular and function approximation settings, and find that with appropriate choices self-consistency can be useful both for policy evaluation and control.
****

Learning on Random Balls is Sufficient for Estimating (Some) Graph Parameters
Takanori Maehara, Hoang NT
Theoretical analyses for graph learning methods often assume a complete observation of the input graph. Such an assumption might not be useful for handling any-size graphs due to the scalability issues in practice. In this work, we develop a theoretical framework for graph classification problems in the partial observation setting (i.e., subgraph samplings). Equipped with insights from graph limit theory, we propose a new graph classification model that works on a randomly sampled subgraph and a novel topology to characterize the representability of the model. Our theoretical framework contributes a theoretical validation of mini-batch learning on graphs and leads to new learning-theoretic results on generalization bounds as well as size-generalizability without assumptions on the input.
****

Risk-Averse Bayes-Adaptive Reinforcement Learning
Marc Rigter, Bruno Lacerda, Nick Hawes
In this work, we address risk-averse Bayes-adaptive reinforcement learning. We pose the problem of optimising the conditional value at risk (CVaR) of the total return in Bayes-adaptive Markov decision processes (MDPs). We show that a policy optimising CVaR in this setting is risk-averse to both the epistemic uncertainty due to the prior distribution over MDPs, and the aleatoric uncertainty due to the inherent stochasticity of MDPs. We reformulate the problem as a two-player stochastic game and propose an approximate algorithm based on Monte Carlo tree search and Bayesian optimisation. Our experiments demonstrate that our approach significantly outperforms baseline approaches for this problem.
****

Iterative Connecting Probability Estimation for Networks
Yichen Qin, Linhan Yu, Yang Li
Estimating the probabilities of connections between vertices in a random network using an observed adjacency matrix is an important task for network data analysis. Many existing estimation methods are based on certain assumptions on network structure, which limit their applicability in practice. Without making strong a

ssumptions, we develop an iterative connecting probability estimation method based on neighborhood averaging. Starting at a random initial point or an existing estimate, our method iteratively updates the pairwise vertex distances, the sets of similar vertices, and connecting probabilities to improve the precision of the estimate. We propose a two-stage neighborhood selection procedure to achieve the trade-off between smoothness of the estimate and the ability to discover local structure. The tuning parameters can be selected by cross-validation. We establish desirable theoretical properties for our method, and further justify its superior performance by comparing with existing methods in simulation and real data analysis.
****
Learning to Adapt via Latent Domains for Adaptive Semantic Segmentation
Yunan Liu, Shanshan Zhang, Yang Li, Jian Yang
Domain adaptive semantic segmentation aims to transfer knowledge learned from labeled source domain to unlabeled target domain. To narrow down the domain gap and ease adaptation difficulty, some recent methods translate source images to target-like images (latent domains), which are used as supplement or substitute to the original source data. Nevertheless, these methods neglect to explicitly model the relationship of knowledge transferring across different domains. Alternatively, in this work we break through the standard “source-target” one pair adaptation framework and construct multiple adaptation pairs (e.g. “source-latent” and “latent-target”). The purpose is to use the meta-knowledge (how to adapt) learned from one pair as guidance to assist the adaptation of another pair under a meta-learning framework. Furthermore, we extend our method to a more practical setting of open compound domain adaptation (a.k.a multiple-target domain adaptation), where the target is a compound of multiple domains without domain labels. In this setting, we embed an additional pair of “latent-latent” to reduce the domain gap between the source and different latent domains, allowing the model to adapt well on multiple target domains simultaneously. When evaluated on standard benchmarks, our method is superior to the state-of-the-art methods in both the single target and multiple-target domain adaptation settings.
****
Single Layer Predictive Normalized Maximum Likelihood for Out-of-Distribution Detection
Koby Bibas, Meir Feder, Tal Hassner
Detecting out-of-distribution (OOD) samples is vital for developing machine learning based models for critical safety systems. Common approaches for OOD detection assume access to some OOD samples during training which may not be available in a real-life scenario. Instead, we utilize the {\em predictive normalized maximum likelihood} (pNML) learner, in which no assumptions are made on the tested input. We derive an explicit expression of the pNML and its generalization error, denoted as the regret, for a single layer neural network (NN). We show that this learner generalizes well when (i) the test vector resides in a subspace spanned by the eigenvectors associated with the large eigenvalues of the empirical correlation matrix of the training data, or (ii) the test sample is far from the decision boundary. Furthermore, we describe how to efficiently apply the derived pNML regret to any pretrained deep NN, by employing the explicit pNML for the last layer, followed by the softmax function. Applying the derived regret to deep NN requires neither additional tunable parameters nor extra data. We extensively evaluate our approach on 74 OOD detection benchmarks using DenseNet-100, ResNet-34, and WideResNet-40 models trained with CIFAR-100, CIFAR-10, SVHN, and ImageNet-30 showing a significant improvement of up to 15.6% over recent leading methods.
****
Prototypical Cross-Attention Networks for Multiple Object Tracking and Segmentation
Lei Ke, Xia Li, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, Fisher Yu
Multiple object tracking and segmentation requires detecting, tracking, and segmenting objects belonging to a set of given classes. Most approaches only exploit the temporal dimension to address the association problem, while relying on sin

gle frame predictions for the segmentation mask itself. We propose Prototypical Cross-Attention Network (PCAN), capable of leveraging rich spatio-temporal information for online multiple object tracking and segmentation. PCAN first distills a space-time memory into a set of prototypes and then employs cross-attention to retrieve rich information from the past frames. To segment each object, PCAN adopts a prototypical appearance module to learn a set of contrastive foreground and background prototypes, which are then propagated over time. Extensive experiments demonstrate that PCAN outperforms current video instance tracking and segmentation competition winners on both Youtube-VIS and BDD100K datasets, and shows efficacy to both one-stage and two-stage segmentation frameworks. Code and video resources are available at http://vis.xyz/pub/pcan.
****
Algorithmic Instabilities of Accelerated Gradient Descent
Amit Attia, Tomer Koren
We study the algorithmic stability of Nesterov's accelerated gradient method. For convex quadratic objectives, Chen et al. (2018) proved that the uniform stability of the method grows quadratically with the number of optimization steps, and conjectured that the same is true for the general convex and smooth case. We disprove this conjecture and show, for two notions of algorithmic stability (including uniform stability), that the stability of Nesterov's accelerated method in fact deteriorates exponentially fast with the number of gradient steps. This stands in sharp contrast to the bounds in the quadratic case, but also to known results for non-accelerated gradient methods where stability typically grows linearly with the number of steps.
****
Learning Optimal Predictive Checklists
Haoran Zhang, Quaid Morris, Berk Ustun, Marzyeh Ghassemi
Checklists are simple decision aids that are often used to promote safety and reliability in clinical applications. In this paper, we present a method to learn checklists for clinical decision support. We represent predictive checklists as discrete linear classifiers with binary features and unit weights. We then learn globally optimal predictive checklists from data by solving an integer programming problem. Our method allows users to customize checklists to obey complex constraints, including constraints to enforce group fairness and to binarize real-valued features at training time. In addition, it pairs models with an optimality gap that can inform model development and determine the feasibility of learning sufficiently accurate checklists on a given dataset. We pair our method with specialized techniques that speed up its ability to train a predictive checklist that performs well and has a small optimality gap. We benchmark the performance of our method on seven clinical classification problems, and demonstrate its practical benefits by training a short-form checklist for PTSD screening. Our results show that our method can fit simple predictive checklists that perform well and that can easily be customized to obey a rich class of custom constraints.
****
Finite Sample Analysis of Average-Reward TD Learning and Q-Learning
Sheng Zhang, Zhe Zhang, Siva Theja Maguluri
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Generalization Bounds for Graph Embedding Using Negative Sampling: Linear vs Hyperbolic
Atsushi Suzuki, Atsushi Nitanda, jing wang, Linchuan Xu, Kenji Yamanishi, Marc Cavazza
Graph embedding, which represents real-world entities in a mathematical space, has enabled numerous applications such as analyzing natural languages, social networks, biochemical networks, and knowledge bases.It has been experimentally shown that graph embedding in hyperbolic space can represent hierarchical tree-like data more effectively than embedding in linear space, owing to hyperbolic space'

s exponential growth property. However, since the theoretical comparison has been limited to ideal noiseless settings, the potential for the hyperbolic space's property to worsen the generalization error for practical data has not been analyzed.In this paper, we provide a generalization error bound applicable for graph embedding both in linear and hyperbolic spaces under various negative sampling settings that appear in graph embedding. Our bound states that error is polynomial and exponential with respect to the embedding space's radius in linear and hyperbolic spaces, respectively, which implies that hyperbolic space's exponential growth property worsens the error.Using our bound, we clarify the data size condition on which graph embedding in hyperbolic space can represent a tree better than in Euclidean space by discussing the bias-variance trade-off.Our bound also shows that imbalanced data distribution, which often appears in graph embedding, can worsen the error.
****
Gradient Starvation: A Learning Proclivity in Neural Networks
Mohammad Pezeshki, Oumar Kaba, Yoshua Bengio, Aaron C. Courville, Doina Precup, Guillaume Lajoie
We identify and formalize a fundamental gradient descent phenomenon resulting in a learning proclivity in over-parameterized neural networks. Gradient Starvation arises when cross-entropy loss is minimized by capturing only a subset of features relevant for the task, despite the presence of other predictive features that fail to be discovered. This work provides a theoretical explanation for the emergence of such feature imbalance in neural networks. Using tools from Dynamical Systems theory, we identify simple properties of learning dynamics during gradient descent that lead to this imbalance, and prove that such a situation can be expected given certain statistical structure in training data. Based on our proposed formalism, we develop guarantees for a novel regularization method aimed at decoupling feature learning dynamics, improving accuracy and robustness in cases hindered by gradient starvation. We illustrate our findings with simple and real-world out-of-distribution (OOD) generalization experiments.
****
Offline Reinforcement Learning as One Big Sequence Modeling Problem
Michael Janner, Qiyang Li, Sergey Levine
Reinforcement learning (RL) is typically viewed as the problem of estimating single-step policies (for model-free RL) or single-step models (for model-based RL), leveraging the Markov property to factorize the problem in time. However, we can also view RL as a sequence modeling problem: predict a sequence of actions that leads to a sequence of high rewards. Viewed in this way, it is tempting to consider whether powerful, high-capacity sequence prediction models that work well in other supervised learning domains, such as natural-language processing, can also provide simple and effective solutions to the RL problem. To this end, we explore how RL can be reframed as "one big sequence modeling" problem, using state-of-the-art Transformer architectures to model distributions over sequences of states, actions, and rewards. Addressing RL as a sequence modeling problem significantly simplifies a range of design decisions: we no longer require separate behavior policy constraints, as is common in prior work on offline model-free RL, and we no longer require ensembles or other epistemic uncertainty estimators, as is common in prior work on model-based RL. All of these roles are filled by the same Transformer sequence model. In our experiments, we demonstrate the flexibility of this approach across imitation learning, goal-conditioned RL, and offline RL.
****
Optimality and Stability in Federated Learning: A Game-theoretic Approach
Kate Donahue, Jon Kleinberg
Federated learning is a distributed learning paradigm where multiple agents, each only with access to local data, jointly learn a global model. There has recently been an explosion of research aiming not only to improve the accuracy rates of federated learning, but also provide certain guarantees around social good properties such as total error. One branch of this research has taken a game-theoretic approach, and in particular, prior work has viewed federated learning as a h

edonic game, where error-minimizing players arrange themselves into federating coalitions. This past work proves the existence of stable coalition partitions, but leaves open a wide range of questions, including how far from optimal these stable solutions are. In this work, we motivate and define a notion of optimality given by the average error rates among federating agents (players). First, we provide and prove the correctness of an efficient algorithm to calculate an optimal (error minimizing) arrangement of players. Next, we analyze the relationship between the stability and optimality of an arrangement. First, we show that for some regions of parameter space, all stable arrangements are optimal (Price of Anarchy equal to 1). However, we show this is not true for all settings: there exist examples of stable arrangements with higher cost than optimal (Price of Anarchy greater than 1). Finally, we give the first constant-factor bound on the performance gap between stability and optimality, proving that the total error of the worst stable solution can be no higher than 9 times the total error of an optimal solution (Price of Anarchy bound of 9).
****

Understanding Deflation Process in Over-parametrized Tensor Decomposition

Rong Ge, Yunwei Ren, Xiang Wang, Mo Zhou

In this paper we study the training dynamics for gradient flow on over-parametrized tensor decomposition problems. Empirically, such training process often first fits larger components and then discovers smaller components, which is similar to a tensor deflation process that is commonly used in tensor decomposition algorithms. We prove that for orthogonally decomposable tensor, a slightly modified version of gradient flow would follow a tensor deflation process and recover all the tensor components. Our proof suggests that for orthogonal tensors, gradient flow dynamics works similarly as greedy low-rank learning in the matrix setting, which is a first step towards understanding the implicit regularization effect of over-parametrized models for low-rank tensors.
****

Privately Learning Subspaces

Vikrant Singhal, Thomas Steinke

Private data analysis suffers a costly curse of dimensionality. However, the data often has an underlying low-dimensional structure. For example, when optimizing via gradient descent, the gradients often lie in or near a low-dimensional subspace. If that low-dimensional structure can be identified, then we can avoid paying (in terms of privacy or accuracy) for the high ambient dimension. We present differentially private algorithms that take input data sampled from a low-dimensional linear subspace (possibly with a small amount of error) and output that subspace (or an approximation to it). These algorithms can serve as a pre-processing step for other procedures.
****

On the Value of Interaction and Function Approximation in Imitation Learning

Nived Rajaraman, Yanjun Han, Lin Yang, Jingbo Liu, 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-authors prior to requesting a name change in the electronic proceedings.
****

Shapeshifter: a Parameter-efficient Transformer using Factorized Reshaped Matrices

Aliakbar Panahi, Seyran Saeedi, Tom Arodz

Language models employ a very large number of trainable parameters. Despite being highly overparameterized, these networks often achieve good out-of-sample test performance on the original task and easily fine-tune to related tasks. Recent observations involving, for example, intrinsic dimension of the objective landscape and the lottery ticket hypothesis, indicate that often training actively involves only a small fraction of the parameter space. Thus, a question remains how large a parameter space needs to be in the first place –- the evidence from recent work on model compression, parameter sharing, factorized representations, a

nd knowledge distillation increasingly shows that models can be made much smalle
r and still perform well. Here, we focus on factorized representations of matric
es that underpin dense, embedding, and self-attention layers. We use low-rank fa
ctorized representation of a reshaped and rearranged original matrix to achieve
space efficient and expressive linear layers. We prove that stacking such low-ra
nk layers increases their expressiveness, providing theoretical understanding fo
r their effectiveness in deep networks. In Transformer models, our approach lead
s to more than ten-fold reduction in the number of total trainable parameters, i
ncluding embedding, attention, and feed-forward layers, with little degradation
in on-task performance. The approach operates out-of-the-box, replacing each pa
rameter matrix with its compact equivalent while maintaining the architecture of
 the network.
****
The Adaptive Doubly Robust Estimator and a Paradox Concerning Logging Policy
Masahiro Kato, Kenichiro McAlinn, Shota Yasui
The doubly robust (DR) estimator, which consists of two nuisance parameters, the
 conditional mean outcome and the logging policy (the probability of choosing an
 action), is crucial in causal inference. This paper proposes a DR estimator for
 dependent samples obtained from adaptive experiments. To obtain an asymptotical
ly normal semiparametric estimator from dependent samples without non-Donsker nu
isance estimators, we propose adaptive-fitting as a variant of sample-splitting.
 We also report an empirical paradox that our proposed DR estimator tends to sho
w better performances compared to other estimators utilizing the true logging po
licy. While a similar phenomenon is known for estimators with i.i.d. samples, tr
aditional explanations based on asymptotic efficiency cannot elucidate our case
with dependent samples. We confirm this hypothesis through simulation studies.
****
Regularized Softmax Deep Multi-Agent Q-Learning
Ling Pan, Tabish Rashid, Bei Peng, Longbo Huang, 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.
****
Physics-Aware Downsampling with Deep Learning for Scalable Flood Modeling
Niv Giladi, Zvika Ben-Haim, Sella Nevo, Yossi Matias, Daniel Soudry
Background. Floods are the most common natural disaster in the world, affecting
the lives of hundreds of millions. Flood forecasting is therefore a vitally impo
rtant endeavor, typically achieved using physical water flow simulations, which
rely on accurate terrain elevation maps. However, such simulations, based on sol
ving partial differential equations, are computationally prohibitive on a large
scale. This scalability issue is commonly alleviated using a coarse grid represe
ntation of the elevation map, though this representation may distort crucial ter
rain details, leading to significant inaccuracies in the simulation.\Contributio
ns. We train a deep neural network to perform physics-informed downsampling of t
he terrain map: we optimize the coarse grid representation of the terrain maps,
so that the flood prediction will match the fine grid solution. For the learning
 process to succeed, we configure a dataset specifically for this task. We demon
strate that with this method, it is possible to achieve a significant reduction
in computational cost, while maintaining an accurate solution. A reference imple
mentation accompanies the paper as well as documentation and code for dataset re
production.
****
Systematic Generalization with Edge Transformers
Leon Bergen, Timothy O'Donnell, Dzmitry Bahdanau
Recent research suggests that systematic generalization in natural language unde
rstanding remains a challenge for state-of-the-art neural models such as Transfo
rmers and Graph Neural Networks. To tackle this challenge, we propose Edge Trans
former, a new model that combines inspiration from Transformers and rule-based s
ymbolic AI. The first key idea in Edge Transformers is to associate vector state

s with every edge, that is, with every pair of input nodes---as opposed to just every node, as it is done in the Transformer model. The second major innovation is a triangular attention mechanism that updates edge representations in a way that is inspired by unification from logic programming. We evaluate Edge Transformer on compositional generalization benchmarks in relational reasoning, semantic parsing, and dependency parsing. In all three settings, the Edge Transformer outperforms Relation-aware, Universal and classical Transformer baselines.
****

TransformerFusion: Monocular RGB Scene Reconstruction using Transformers
Aljaz Bozic, Pablo Palafox, Justus Thies, Angela Dai, Matthias Niessner
We introduce TransformerFusion, a transformer-based 3D scene reconstruction approach. From an input monocular RGB video, the video frames are processed by a transformer network that fuses the observations into a volumetric feature grid representing the scene; this feature grid is then decoded into an implicit 3D scene representation. Key to our approach is the transformer architecture that enables the network to learn to attend to the most relevant image frames for each 3D location in the scene, supervised only by the scene reconstruction task. Features are fused in a coarse-to-fine fashion, storing fine-level features only where needed, requiring lower memory storage and enabling fusion at interactive rates. The feature grid is then decoded to a higher-resolution scene reconstruction, using an MLP-based surface occupancy prediction from interpolated coarse-to-fine 3D features. Our approach results in an accurate surface reconstruction, outperforming state-of-the-art multi-view stereo depth estimation methods, fully-convolutional 3D reconstruction approaches, and approaches using LSTM- or GRU-based recurrent networks for video sequence fusion.
****

Maximum Likelihood Training of Score-Based Diffusion Models
Yang Song, Conor Durkan, Iain Murray, Stefano Ermon
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

Global Convergence of Gradient Descent for Asymmetric Low-Rank Matrix Factorization
Tian Ye, Simon S. Du
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 Data Augmentation on Temporal Graphs
Yiwei Wang, Yujun Cai, Yuxuan Liang, Henghui Ding, Changhu Wang, Siddharth Bhatia, Bryan Hooi
Temporal Graph Networks (TGNs) are powerful on modeling temporal graph data based on their increased complexity. Higher complexity carries with it a higher risk of overfitting, which makes TGNs capture random noise instead of essential semantic information. To address this issue, our idea is to transform the temporal graphs using data augmentation (DA) with adaptive magnitudes, so as to effectively augment the input features and preserve the essential semantic information. Based on this idea, we present the MeTA (Memory Tower Augmentation) module: a multi-level module that processes the augmented graphs of different magnitudes on separate levels, and performs message passing across levels to provide adaptively augmented inputs for every prediction. MeTA can be flexibly applied to the training of popular TGNs to improve their effectiveness without increasing their time complexity. To complement MeTA, we propose three DA strategies to realistically model noise by modifying both the temporal and topological features. Empirical results on standard datasets show that MeTA yields significant gains for the popular TGN models on edge prediction and node classification in an efficient manner.

****
Regularized Frank-Wolfe for Dense CRFs: Generalizing Mean Field and Beyond
■.Khuê Lê-Huu, Karteek Alahari
We introduce regularized Frank-Wolfe, a general and effective algorithm for infe
rence and learning of dense conditional random fields (CRFs). The algorithm opti
mizes a nonconvex continuous relaxation of the CRF inference problem using vanil
la Frank-Wolfe with approximate updates, which are equivalent to minimizing a re
gularized energy function. Our proposed method is a generalization of existing a
lgorithms such as mean field or concave-convex procedure. This perspective not o
nly offers a unified analysis of these algorithms, but also allows an easy way o
f exploring different variants that potentially yield better performance. We ill
ustrate this in our empirical results on standard semantic segmentation datasets
, where several instantiations of our regularized Frank-Wolfe outperform mean fi
eld inference, both as a standalone component and as an end-to-end trainable lay
er in a neural network. We also show that dense CRFs, coupled with our new algor
ithms, produce significant improvements over strong CNN baselines.
****
Terra: Imperative-Symbolic Co-Execution of Imperative Deep Learning Programs
Taebum Kim, Eunji Jeong, Geon-Woo Kim, Yunmo Koo, Sehoon Kim, Gyeongin Yu, Byung
-Gon Chun
Imperative programming allows users to implement their deep neural networks (DNN
s) easily and has become an essential part of recent deep learning (DL) framewor
ks. Recently, several systems have been proposed to combine the usability of imp
erative programming with the optimized performance of symbolic graph execution.
Such systems convert imperative Python DL programs to optimized symbolic graphs
and execute them. However, they cannot fully support the usability of imperative
 programming. For example, if an imperative DL program contains a Python feature
 with no corresponding symbolic representation (e.g., third-party library calls
or unsupported dynamic control flows) they fail to execute the program. To overc
ome this limitation, we propose Terra, an imperative-symbolic co-execution syste
m that can handle any imperative DL programs while achieving the optimized perfo
rmance of symbolic graph execution. To achieve this, Terra builds a symbolic gra
ph by decoupling DL operations from Python features. Then, Terra conducts the im
perative execution to support all Python features, while delegating the decouple
d operations to the symbolic execution. We evaluated Terra’s performance improve
ment and coverage with ten imperative DL programs for several DNN architectures.
 The results show that Terra can speed up the execution of all ten imperative DL
 programs, whereas AutoGraph, one of the state-of-the-art systems, fails to exec
ute five of them.
****
Uniform Sampling over Episode Difficulty
Sébastien Arnold, Guneet Dhillon, Avinash Ravichandran, Stefano Soatto
Episodic training is a core ingredient of few-shot learning to train models on t
asks with limited labelled data. Despite its success, episodic training remains
largely understudied, prompting us to ask the question: what is the best way to
sample episodes? In this paper, we first propose a method to approximate episode
 sampling distributions based on their difficulty. Building on this method, we p
erform an extensive analysis and find that sampling uniformly over episode diffi
culty outperforms other sampling schemes, including curriculum and easy-/hard-mi
ning. As the proposed sampling method is algorithm agnostic, we can leverage the
se insights to improve few-shot learning accuracies across many episodic trainin
g algorithms. We demonstrate the efficacy of our method across popular few-shot
learning datasets, algorithms, network architectures, and protocols.
****
Scalable Intervention Target Estimation in Linear Models
Burak Varici, Karthikeyan Shanmugam, Prasanna Sattigeri, Ali Tajer
This paper considers the problem of estimating the unknown intervention targets
in a causal directed acyclic graph from observational and interventional data. T
he focus is on soft interventions in linear structural equation models (SEMs). C
urrent approaches to causal structure learning either work with known interventi

on targets or use hypothesis testing to discover the unknown intervention target
s even for linear SEMs. This severely limits their scalability and sample comple
xity. This paper proposes a scalable and efficient algorithm that consistently i
dentifies all intervention targets. The pivotal idea is to estimate the interven
tion sites from the difference between the precision matrices associated with th
e observational and interventional datasets. It involves repeatedly estimating s
uch sites in different subsets of variables. The proposed algorithm can be used
to also update a given observational Markov equivalence class into the intervent
ional Markov equivalence class. Consistency, Markov equivalency, and sample comp
lexity are established analytically. Finally, simulation results on both real an
d synthetic data demonstrate the gains of the proposed approach for scalable cau
sal structure recovery. Implementation of the algorithm and the code to reproduc
e the simulation results are available at \url{https://github.com/bvarici/interv
ention-estimation}.
****
Play to Grade: Testing Coding Games as Classifying Markov Decision Process
Allen Nie, Emma Brunskill, Chris Piech
Contemporary coding education often presents students with the task of developin
g programs that have user interaction and complex dynamic systems, such as mouse
 based games. While pedagogically compelling, there are no contemporary autonomo
us methods for providing feedback. Notably, interactive programs are impossible
to grade by traditional unit tests. In this paper we formalize the challenge of
providing feedback to interactive programs as a task of classifying Markov Decis
ion Processes (MDPs). Each student's program fully specifies an MDP where the ag
ent needs to operate and decide, under reasonable generalization, if the dynamic
s and reward model of the input MDP should be categorized as correct or broken.
We demonstrate that by designing a cooperative objective between an agent and an
 autoregressive model, we can use the agent to sample differential trajectories
from the input MDP that allows a classifier to determine membership: Play to Gra
de. Our method enables an automatic feedback system for interactive code assignm
ents. We release a dataset of 711,274 anonymized student submissions to a single
 assignment with hand-coded bug labels to support future research.
****
Distributional Reinforcement Learning for Multi-Dimensional Reward Functions
Pushi Zhang, Xiaoyu Chen, Li Zhao, Wei Xiong, Tao Qin, Tie-Yan Liu
A growing trend for value-based reinforcement learning (RL) algorithms is to cap
ture more information than scalar value functions in the value network. One of t
he most well-known methods in this branch is distributional RL, which models ret
urn distribution instead of scalar value. In another line of work, hybrid reward
 architectures (HRA) in RL have studied to model source-specific value functions
 for each source of reward, which is also shown to be beneficial in performance.
 To fully inherit the benefits of distributional RL and hybrid reward architectu
res, we introduce Multi-Dimensional Distributional DQN (MD3QN), which extends di
stributional RL to model the joint return distribution from multiple reward sour
ces. As a by-product of joint distribution modeling, MD3QN can capture not only
the randomness in returns for each source of reward, but also the rich reward co
rrelation between the randomness of different sources. We prove the convergence
for the joint distributional Bellman operator and build our empirical algorithm
by minimizing the Maximum Mean Discrepancy between joint return distribution and
 its Bellman target. In experiments, our method accurately models the joint retu
rn distribution in environments with richly correlated reward functions, and out
performs previous RL methods utilizing multi-dimensional reward functions in the
 control setting.
****
Differentiable Unsupervised Feature Selection based on a Gated Laplacian
Ofir Lindenbaum, Uri Shaham, Erez Peterfreund, Jonathan Svirsky, Nicolas Casey,
Yuval Kluger
Scientific observations may consist of a large number of variables (features). S
electing a subset of meaningful features is often crucial for identifying patter
ns hidden in the ambient space. In this paper, we present a method for unsupervi

sed feature selection, and we demonstrate its advantage in clustering, a common unsupervised task. We propose a differentiable loss that combines a graph Laplacian-based score that favors low-frequency features with a gating mechanism for removing nuisance features. Our method improves upon the naive graph Laplacian score by replacing it with a gated variant computed on a subset of low-frequency features. We identify this subset by learning the parameters of continuously relaxed Bernoulli variables, which gate the entire feature space. We mathematically motivate the proposed approach and demonstrate that it is crucial to compute the graph Laplacian on the gated inputs rather than on the full feature set in the high noise regime. Using several real-world examples, we demonstrate the efficacy and advantage of the proposed approach over leading baselines.
****
Smooth Bilevel Programming for Sparse Regularization
Clarice Poon, Gabriel Peyré
Iteratively reweighted least square (IRLS) is a popular approach to solve sparsity-enforcing regression problems in machine learning. State of the art approaches are more efficient but typically rely on specific coordinate pruning schemes. In this work, we show how a surprisingly simple re-parametrization of IRLS, coupled with a bilevel resolution (instead of an alternating scheme) is able to achieve top performances on a wide range of sparsity (such as Lasso, group Lasso and trace norm regularizations), regularization strength (including hard constraints), and design matrices (ranging from correlated designs to differential operators). Similarly to IRLS, our method only involves linear systems resolutions, but in sharp contrast, corresponds to the minimization of a smooth function. Despite being non-convex, we show that there is no spurious minima and that saddle points are "ridable'', so that there always exists a descent direction. We thus advocate for the use of a BFGS quasi-Newton solver, which makes our approach simple, robust and efficient. We perform a numerical benchmark of the convergence speed of our algorithm against state of the art solvers for Lasso, group Lasso, trace norm and linearly constrained problems. These results highlight the versatility of our approach, removing the need to use different solvers depending on the specificity of the ML problem under study.
****
Grounding Representation Similarity Through Statistical Testing
Frances Ding, Jean-Stanislas Denain, Jacob Steinhardt
To understand neural network behavior, recent works quantitatively compare different networks' learned representations using canonical correlation analysis (CCA), centered kernel alignment (CKA), and other dissimilarity measures. Unfortunately, these widely used measures often disagree on fundamental observations, such as whether deep networks differing only in random initialization learn similar representations. These disagreements raise the question: which, if any, of these dissimilarity measures should we believe? We provide a framework to ground this question through a concrete test: measures should have \emph{sensitivity} to changes that affect functional behavior, and \emph{specificity} against changes that do not. We quantify this through a variety of functional behaviors including probing accuracy and robustness to distribution shift, and examine changes such as varying random initialization and deleting principal components. We find that current metrics exhibit different weaknesses, note that a classical baseline performs surprisingly well, and highlight settings where all metrics appear to fail, thus providing a challenge set for further improvement.
****
A Consciousness-Inspired Planning Agent for Model-Based Reinforcement Learning
Mingde Zhao, Zhen Liu, Sitao Luan, Shuyuan Zhang, Doina Precup, Yoshua Bengio
We present an end-to-end, model-based deep reinforcement learning agent which dynamically attends to relevant parts of its state during planning. The agent uses a bottleneck mechanism over a set-based representation to force the number of entities to which the agent attends at each planning step to be small. In experiments, we investigate the bottleneck mechanism with several sets of customized environments featuring different challenges. We consistently observe that the design allows the planning agents to generalize their learned task-solving abilities

in compatible unseen environments by attending to the relevant objects, leading to better out-of-distribution generalization performance.
****
Reward-Free Model-Based Reinforcement Learning with Linear Function Approximation
Weitong ZHANG, Dongruo Zhou, 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.
****
Beltrami Flow and Neural Diffusion on Graphs
Benjamin Chamberlain, James Rowbottom, Davide Eynard, Francesco Di Giovanni, Xiaowen Dong, Michael Bronstein
We propose a novel class of graph neural networks based on the discretized Beltrami flow, a non-Euclidean diffusion PDE. In our model, node features are supplemented with positional encodings derived from the graph topology and jointly evolved by the Beltrami flow, producing simultaneously continuous feature learning, topology evolution. The resulting model generalizes many popular graph neural networks and achieves state-of-the-art results on several benchmarks.
****
Think Big, Teach Small: Do Language Models Distil Occam’s Razor?
Gonzalo Jaimovitch-Lopez, David Castellano Falcón, Cesar Ferri, José Hernández-Orallo
Large language models have recently shown a remarkable ability for few-shot learning, including patterns of algorithmic nature. However, it is still an open question to determine what kind of patterns these models can capture and how many examples they need in their prompts. We frame this question as a teaching problem with strong priors, and study whether language models can identify simple algorithmic concepts from small witness sets. In particular, we explore how several GPT architectures, program induction systems and humans perform in terms of the complexity of the concept and the number of additional examples, and how much their behaviour differs. This first joint analysis of language models and machine teaching can address key questions for artificial intelligence and machine learning, such as whether some strong priors, and Occam’s razor in particular, can be distilled from data, making learning from a few examples possible.
****
Disentangling Identifiable Features from Noisy Data with Structured Nonlinear ICA
Hermanni Hälvä, Sylvain Le Corff, Luc Lehéricy, Jonathan So, Yongjie Zhu, Elisabeth Gassiat, Aapo Hyvarinen
We introduce a new general identifiable framework for principled disentanglement referred to as Structured Nonlinear Independent Component Analysis (SNICA). Our contribution is to extend the identifiability theory of deep generative models for a very broad class of structured models. While previous works have shown identifiability for specific classes of time-series models, our theorems extend this to more general temporal structures as well as to models with more complex structures such as spatial dependencies. In particular, we establish the major result that identifiability for this framework holds even in the presence of noise of unknown distribution. Finally, as an example of our framework's flexibility, we introduce the first nonlinear ICA model for time-series that combines the following very useful properties: it accounts for both nonstationarity and autocorrelation in a fully unsupervised setting; performs dimensionality reduction; models hidden states; and enables principled estimation and inference by variational maximum-likelihood.
****
Conditionally Parameterized, Discretization-Aware Neural Networks for Mesh-Based Modeling of Physical Systems
Jiayang Xu, Aniruddhe Pradhan, Karthikeyan Duraisamy
Simulations of complex physical systems are typically realized by discretizing p

artial differential equations (PDEs) on unstructured meshes. While neural networks have recently been explored for the surrogate and reduced order modeling of PDE solutions, they often ignore interactions or hierarchical relations between input features, and process them as concatenated mixtures. We generalize the idea of conditional parameterization -- using trainable functions of input parameters to generate the weights of a neural network, and extend them in a flexible way to encode critical information. Inspired by discretized numerical methods, choices of the parameters include physical quantities and mesh topology features. The functional relation between the modeled features and the parameters is built into the network architecture. The method is implemented on different networks and applied to frontier scientific machine learning tasks including the discovery of unmodeled physics, super-resolution of coarse fields, and the simulation of unsteady flows with chemical reactions. The results show that the conditionally-parameterized networks provide superior performance compared to their traditional counterparts. The CP-GNet - an architecture that can be trained on very few data snapshots - is proposed as the first deep learning model capable of standalone prediction of reacting flows on irregular meshes.
****
USCO-Solver: Solving Undetermined Stochastic Combinatorial Optimization Problems
Guangmo Tong
Real-world decision-making systems are often subject to uncertainties that have to be resolved through observational data. Therefore, we are frequently confronted with combinatorial optimization problems of which the objective function is unknown and thus has to be debunked using empirical evidence. In contrast to the common practice that relies on a learning-and-optimization strategy, we consider the regression between combinatorial spaces, aiming to infer high-quality optimization solutions from samples of input-solution pairs -- without the need to learn the objective function. Our main deliverable is a universal solver that is able to handle abstract undetermined stochastic combinatorial optimization problems. For learning foundations, we present learning-error analysis under the PAC-Bayesian framework using a new margin-based analysis. In empirical studies, we demonstrate our design using proof-of-concept experiments, and compare it with other methods that are potentially applicable. Overall, we obtain highly encouraging experimental results for several classic combinatorial problems on both synthetic and real-world datasets.
****
Adaptive Conformal Inference Under Distribution Shift
Isaac Gibbs, Emmanuel Candes
We develop methods for forming prediction sets in an online setting where the data generating distribution is allowed to vary over time in an unknown fashion. Our framework builds on ideas from conformal inference to provide a general wrapper that can be combined with any black box method that produces point predictions of the unseen label or estimated quantiles of its distribution. While previous conformal inference methods rely on the assumption that the data are exchangeable, our adaptive approach provably achieves the desired coverage frequency over long-time intervals irrespective of the true data generating process. We accomplish this by modelling the distribution shift as a learning problem in a single parameter whose optimal value is varying over time and must be continuously re-estimated. We test our method, adaptive conformal inference, on two real world datasets and find that its predictions are robust to visible and significant distribution shifts.
****
Periodic Activation Functions Induce Stationarity
Lassi Meronen, Martin Trapp, Arno Solin
Neural network models are known to reinforce hidden data biases, making them unreliable and difficult to interpret. We seek to build models that `know what they do not know' by introducing inductive biases in the function space. We show that periodic activation functions in Bayesian neural networks establish a connection between the prior on the network weights and translation-invariant, stationary Gaussian process priors. Furthermore, we show that this link goes beyond sinus

oidal (Fourier) activations by also covering triangular wave and periodic ReLU a
ctivation functions. In a series of experiments, we show that periodic activatio
n functions obtain comparable performance for in-domain data and capture sensiti
vity to perturbed inputs in deep neural networks for out-of-domain detection.
****
Towards Optimal Strategies for Training Self-Driving Perception Models in Simula
tion
David Acuna, Jonah Philion, Sanja Fidler
Autonomous driving relies on a huge volume of real-world data to be labeled to h
igh precision. Alternative solutions seek to exploit driving simulators that ca
n generate large amounts of labeled data with a plethora of content variations.
However, the domain gap between the synthetic and real data remains, raising th
e following important question: What are the best way to utilize a self-driving
simulator for perception tasks?. In this work, we build on top of recent advance
s in domain-adaptation theory, and from this perspective, propose ways to minimi
ze the reality gap. We primarily focus on the use of labels in the synthetic dom
ain alone. Our approach introduces both a principled way to learn neural-invaria
nt representations and a theoretically inspired view on how to sample the data
from the simulator. Our method is easy to implement in practice as it is agnosti
c of the network architecture and the choice of the simulator.   We showcase our
 approach on the bird's-eye-view vehicle segmentation task with multi-sensor dat
a (cameras, lidar) using an open-source simulator (CARLA), and evaluate the enti
re framework on a real-world dataset (nuScenes). Last but not least, we show wha
t types of variations (e.g. weather conditions, number of assets, map design and
 color diversity) matter to perception networks when trained with driving simula
tors, and which ones can be compensated for with our domain adaptation technique
.
****
KS-GNN: Keywords Search over Incomplete Graphs via Graphs Neural Network
YU HAO, Xin Cao, Yufan Sheng, Yixiang Fang, Wei Wang
Keyword search is a fundamental task to retrieve information that is the most re
levant to the query keywords. Keyword search over graphs aims to find subtrees o
r subgraphs containing all query keywords ranked according to some criteria. Exi
sting studies all assume that the graphs have complete information. However, rea
l-world graphs may contain some missing information (such as edges or keywords),
 thus making the problem much more challenging. To solve the problem of keyword
search over incomplete graphs, we propose a novel model named KS-GNN based on th
e graph neural network and the auto-encoder. By considering the latent relations
hips and the frequency of different keywords, the proposed KS-GNN aims to allevi
ate the effect of missing information and is able to learn low-dimensional repre
sentative node embeddings that preserve both graph structure and keyword feature
s. Our model can effectively answer keyword search queries with linear time comp
lexity over incomplete graphs. The experiments on four real-world datasets show
that our model consistently achieves better performance than state-of-the-art ba
seline methods in graphs having missing information.
****
Reconstruction for Powerful Graph Representations
Leonardo Cotta, Christopher Morris, Bruno Ribeiro
Graph neural networks (GNNs) have limited expressive power, failing to represent
 many graph classes correctly. While more expressive graph representation learni
ng (GRL) alternatives can distinguish some of these classes, they are significan
tly harder to implement, may not scale well, and have not been shown to outperfo
rm well-tuned GNNs in real-world tasks. Thus, devising simple, scalable, and exp
ressive GRL architectures that also achieve real-world improvements remains an o
pen challenge. In this work, we show the extent to which graph reconstruction---
reconstructing a graph from its subgraphs---can mitigate the theoretical and pra
ctical problems currently faced by GRL architectures. First, we leverage graph r
econstruction to build two new classes of expressive graph representations. Seco
ndly, we show how graph reconstruction boosts the expressive power of any GNN ar
chitecture while being a (provably) powerful inductive bias for invariances to v

ertex removals. Empirically, we show how reconstruction can boost GNN's expressive power---while maintaining its invariance to permutations of the vertices---by solving seven graph property tasks not solvable by the original GNN. Further, we demonstrate how it boosts state-of-the-art GNN's performance across nine real-world benchmark datasets.
****
Revealing and Protecting Labels in Distributed Training
Trung Dang, Om Thakkar, Swaroop Ramaswamy, Rajiv Mathews, Peter Chin, Françoise Beaufays
Distributed learning paradigms such as federated learning often involve transmission of model updates, or gradients, over a network, thereby avoiding transmission of private data. However, it is possible for sensitive information about the training data to be revealed from such gradients. Prior works have demonstrated that labels can be revealed analytically from the last layer of certain models (e.g., ResNet), or they can be reconstructed jointly with model inputs by using Gradients Matching [Zhu et al.] with additional knowledge about the current state of the model. In this work, we propose a method to discover the set of labels of training samples from only the gradient of the last layer and the id to label mapping. Our method is applicable to a wide variety of model architectures across multiple domains. We demonstrate the effectiveness of our method for model training in two domains - image classification, and automatic speech recognition. Furthermore, we show that existing reconstruction techniques improve their efficacy when used in conjunction with our method. Conversely, we demonstrate that gradient quantization and sparsification can significantly reduce the success of the attack.
****
Solving Graph-based Public Goods Games with Tree Search and Imitation Learning
Victor-Alexandru Darvariu, Stephen Hailes, Mirco Musolesi
Public goods games represent insightful settings for studying incentives for individual agents to make contributions that, while costly for each of them, benefit the wider society. In this work, we adopt the perspective of a central planner with a global view of a network of self-interested agents and the goal of maximizing some desired property in the context of a best-shot public goods game. Existing algorithms for this known NP-complete problem find solutions that are sub-optimal and cannot optimize for criteria other than social welfare.In order to efficiently solve public goods games, our proposed method directly exploits the correspondence between equilibria and the Maximal Independent Set (mIS) structural property of graphs. In particular, we define a Markov Decision Process which incrementally generates an mIS, and adopt a planning method to search for equilibria, outperforming existing methods. Furthermore, we devise a graph imitation learning technique that uses demonstrations of the search to obtain a graph neural network parametrized policy which quickly generalizes to unseen game instances. Our evaluation results show that this policy is able to reach 99.5\% of the performance of the planning method while being three orders of magnitude faster to evaluate on the largest graphs tested. The methods presented in this work can be applied to a large class of public goods games of potentially high societal impact and more broadly to other graph combinatorial optimization problems.
****
Stochastic Optimization of Areas Under Precision-Recall Curves with Provable Convergence
Qi Qi, Youzhi Luo, Zhao Xu, Shuiwang Ji, Tianbao Yang
Areas under ROC (AUROC) and precision-recall curves (AUPRC) are common metrics for evaluating classification performance for imbalanced problems. Compared with AUROC, AUPRC is a more appropriate metric for highly imbalanced datasets. While stochastic optimization of AUROC has been studied extensively, principled stochastic optimization of AUPRC has been rarely explored. In this work, we propose a principled technical method to optimize AUPRC for deep learning. Our approach is based on maximizing the averaged precision (AP), which is an unbiased point estimator of AUPRC. We cast the objective into a sum of dependent compositional functions with inner functions dependent on random variables of the outer level. We

propose efficient adaptive and non-adaptive stochastic algorithms named SOAP with provable convergence guarantee under mild conditions by leveraging recent advances in stochastic compositional optimization. Extensive experimental results on image and graph datasets demonstrate that our proposed method outperforms prior methods on imbalanced problems in terms of AUPRC. To the best of our knowledge, our work represents the first attempt to optimize AUPRC with provable convergence. The SOAP has been implemented in the libAUC library at https://libauc.org/.
****
Transfer Learning of Graph Neural Networks with Ego-graph Information Maximization
Qi Zhu, Carl Yang, Yidan Xu, Haonan Wang, Chao Zhang, Jiawei Han
Graph neural networks (GNNs) have achieved superior performance in various applications, but training dedicated GNNs can be costly for large-scale graphs. Some recent work started to study the pre-training of GNNs. However, none of them provide theoretical insights into the design of their frameworks, or clear requirements and guarantees towards their transferability. In this work, we establish a theoretically grounded and practically useful framework for the transfer learning of GNNs. Firstly, we propose a novel view towards the essential graph information and advocate the capturing of it as the goal of transferable GNN training, which motivates the design of EGI (Ego-Graph Information maximization) to analytically achieve this goal. Secondly,when node features are structure-relevant, we conduct an analysis of EGI transferability regarding the difference between the local graph Laplacians of the source and target graphs. We conduct controlled synthetic experiments to directly justify our theoretical conclusions. Comprehensive experiments on two real-world network datasets show consistent results in the analyzed setting of direct-transfering, while those on large-scale knowledge graphs show promising results in the more practical setting of transfering with fine-tuning.
****
You are caught stealing my winning lottery ticket! Making a lottery ticket claim its ownership
Xuxi Chen, Tianlong Chen, Zhenyu Zhang, Zhangyang 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.
****
Complexity Lower Bounds for Nonconvex-Strongly-Concave Min-Max Optimization
Haochuan Li, Yi Tian, Jingzhao Zhang, Ali Jadbabaie
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Early-stopped neural networks are consistent
Ziwei Ji, Justin Li, Matus Telgarsky
This work studies the behavior of shallow ReLU networks trained with the logistic loss via gradient descent on binary classification data where the underlying data distribution is general, and the (optimal) Bayes risk is not necessarily zero. In this setting, it is shown that gradient descent with early stopping achieves population risk arbitrarily close to optimal in terms of not just logistic and misclassification losses, but also in terms of calibration, meaning the sigmoid mapping of its outputs approximates the true underlying conditional distribution arbitrarily finely. Moreover, the necessary iteration, sample, and architectural complexities of this analysis all scale naturally with a certain complexity measure of the true conditional model. Lastly, while it is not shown that early stopping is necessary, it is shown that any classifier satisfying a basic local interpolation property is inconsistent.
****
NxMTransformer: Semi-Structured Sparsification for Natural Language Understandin

g via ADMM
Connor Holmes, Minjia Zhang, Yuxiong He, Bo Wu
Natural Language Processing (NLP) has recently achieved great success by using huge pre-trained Transformer networks. However, these models often contain hundreds of millions or even billions of parameters, bringing challenges to online deployment due to latency constraints. Recently, hardware manufacturers have introduced dedicated hardware for NxM sparsity to provide the flexibility of unstructured pruning with the runtime efficiency of structured approaches. NxM sparsity permits arbitrarily selecting M parameters to retain from a contiguous group of N in the dense representation. However, due to the extremely high complexity of pre-trained models, the standard sparse fine-tuning techniques often fail to generalize well on downstream tasks, which have limited data resources. To address such an issue in a principled manner, we introduce a new learning framework, called NxMTransformer, to induce NxM semi-structured sparsity on pretrained language models for natural language understanding to obtain better performance. In particular, we propose to formulate the NxM sparsity as a constrained optimization problem and use Alternating Direction Method of Multipliers (ADMM) to optimize the downstream tasks while taking the underlying hardware constraints into consideration. ADMM decomposes the NxM sparsification problem into two sub-problems that can be solved sequentially, generating sparsified Transformer networks that achieve high accuracy while being able to effectively execute on newly released hardware. We apply our approach to a wide range of NLP tasks, and our proposed method is able to achieve 1.7 points higher accuracy in GLUE score than current best practices. Moreover, we perform detailed analysis on our approach and shed light on how ADMM affects fine-tuning accuracy for downstream tasks. Finally, we illustrate how NxMTransformer achieves additional performance improvement with knowledge distillation based methods.
****
Reliable Decisions with Threshold Calibration
Roshni Sahoo, Shengjia Zhao, Alyssa Chen, Stefano Ermon
Decision makers rely on probabilistic forecasts to predict the loss of different decision rules before deployment. When the forecasted probabilities match the true frequencies, predicted losses will be accurate. Although perfect forecasts are typically impossible, probabilities can be calibrated to match the true frequencies on average. However, we find that this \textit{average} notion of calibration, which is typically used in practice, does not necessarily guarantee accurate decision loss prediction. Specifically in the regression setting, the loss of threshold decisions, which are decisions based on whether the forecasted outcome falls above or below a cutoff, might not be predicted accurately. We propose a stronger notion of calibration called threshold calibration, which is exactly the condition required to ensure that decision loss is predicted accurately for threshold decisions. We provide an efficient algorithm which takes an uncalibrated forecaster as input and provably outputs a threshold-calibrated forecaster. Our procedure allows downstream decision makers to confidently estimate the loss of any threshold decision under any threshold loss function. Empirically, threshold calibration improves decision loss prediction without compromising on the quality of the decisions in two real-world settings: hospital scheduling decisions and resource allocation decisions.
****
End-to-End Weak Supervision
Salva Rühling Cachay, Benedikt Boecking, Artur Dubrawski
Aggregating multiple sources of weak supervision (WS) can ease the data-labeling bottleneck prevalent in many machine learning applications, by replacing the tedious manual collection of ground truth labels. Current state of the art approaches that do not use any labeled training data, however, require two separate modeling steps: Learning a probabilistic latent variable model based on the WS sources -- making assumptions that rarely hold in practice -- followed by downstream model training. Importantly, the first step of modeling does not consider the performance of the downstream model.To address these caveats we propose an end-to-end approach for directly learning the downstream model by maximizing its agree

ment with probabilistic labels generated by reparameterizing previous probabilistic posteriors with a neural network. Our results show improved performance over prior work in terms of end model performance on downstream test sets, as well as in terms of improved robustness to dependencies among weak supervision sources.
****
Shift Invariance Can Reduce Adversarial Robustness
Vasu Singla, Songwei Ge, Basri Ronen, David Jacobs
Shift invariance is a critical property of CNNs that improves performance on classification. However, we show that invariance to circular shifts can also lead to greater sensitivity to adversarial attacks. We first characterize the margin between classes when a shift-invariant {\em linear} classifier is used. We show that the margin can only depend on the DC component of the signals. Then, using results about infinitely wide networks, we show that in some simple cases, fully connected and shift-invariant neural networks produce linear decision boundaries. Using this, we prove that shift invariance in neural networks produces adversarial examples for the simple case of two classes, each consisting of a single image with a black or white dot on a gray background. This is more than a curiosity; we show empirically that with real datasets and realistic architectures, shift invariance reduces adversarial robustness. Finally, we describe initial experiments using synthetic data to probe the source of this connection.
****
Wisdom of the Crowd Voting: Truthful Aggregation of Voter Information and Preferences
Grant Schoenebeck, Biaoshuai Tao
We consider two-alternative elections where voters' preferences depend on a state variable that is not directly observable. Each voter receives a private signal that is correlated to the state variable. As a special case, our model captures the common scenario where voters can be categorized into three types: those who always prefer one alternative, those who always prefer the other, and those contingent voters whose preferences depends on the state. In this setting, even if every voter is a contingent voter, agents voting according to their private information need not result in the adoption of the universally preferred alternative, because the signals can be systematically biased.We present a mechanism that elicits and aggregates the private signals from the voters, and outputs the alternative that is favored by the majority. In particular, voters truthfully reporting their signals forms a strong Bayes Nash equilibrium (where no coalition of voters can deviate and receive a better outcome).
****
Replay-Guided Adversarial Environment Design
Minqi Jiang, Michael Dennis, Jack Parker-Holder, Jakob Foerster, Edward Grefenstette, Tim Rocktäschel
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
There Is No Turning Back: A Self-Supervised Approach for Reversibility-Aware Reinforcement Learning
Nathan Grinsztajn, Johan Ferret, Olivier Pietquin, philippe preux, Matthieu Geist
We propose to learn to distinguish reversible from irreversible actions for better informed decision-making in Reinforcement Learning (RL). From theoretical considerations, we show that approximate reversibility can be learned through a simple surrogate task: ranking randomly sampled trajectory events in chronological order. Intuitively, pairs of events that are always observed in the same order are likely to be separated by an irreversible sequence of actions. Conveniently, learning the temporal order of events can be done in a fully self-supervised way, which we use to estimate the reversibility of actions from experience, without any priors.We propose two different strategies that incorporate reversibility i

n RL agents, one strategy for exploration (RAE) and one strategy for control (RA C). We demonstrate the potential of reversibility-aware agents in several enviro nments, including the challenging Sokoban game. In synthetic tasks, we show that we can learn control policies that never fail and reduce to zero the side-effec ts of interactions, even without access to the reward function.
****
Learning to Execute: Efficient Learning of Universal Plan-Conditioned Policies i n Robotics
Ingmar Schubert, Danny Driess, Ozgur S. Oguz, Marc Toussaint
Applications of Reinforcement Learning (RL) in robotics are often limited by hig h data demand. On the other hand, approximate models are readily available in ma ny robotics scenarios, making model-based approaches like planning a data-effici ent alternative. Still, the performance of these methods suffers if the model is imprecise or wrong. In this sense, the respective strengths and weaknesses of R L and model-based planners are complementary. In the present work, we investigat e how both approaches can be integrated into one framework that combines their s trengths. We introduce Learning to Execute (L2E), which leverages information co ntained in approximate plans to learn universal policies that are conditioned on plans. In our robotic manipulation experiments, L2E exhibits increased performa nce when compared to pure RL, pure planning, or baseline methods combining learn ing and planning.
****
Self-Diagnosing GAN: Diagnosing Underrepresented Samples in Generative Adversari al Networks
Jinhee Lee, Haeri Kim, Youngkyu Hong, Hye Won Chung
Despite remarkable performance in producing realistic samples, Generative Advers arial Networks (GANs) often produce low-quality samples near low-density regions of the data manifold, e.g., samples of minor groups. Many techniques have been developed to improve the quality of generated samples, either by post-processing generated samples or by pre-processing the empirical data distribution, but at the cost of reduced diversity. To promote diversity in sample generation without degrading the overall quality, we propose a simple yet effective method to diag nose and emphasize underrepresented samples during training of a GAN. The main i dea is to use the statistics of the discrepancy between the data distribution an d the model distribution at each data instance. Based on the observation that th e underrepresented samples have a high average discrepancy or high variability i n discrepancy, we propose a method to emphasize those samples during training of a GAN. Our experimental results demonstrate that the proposed method improves G AN performance on various datasets, and it is especially effective in improving the quality and diversity of sample generation for minor groups.
****
Online Multi-Armed Bandits with Adaptive Inference
Maria Dimakopoulou, Zhimei Ren, Zhengyuan Zhou
During online decision making in Multi-Armed Bandits (MAB), one needs to conduct inference on the true mean reward of each arm based on data collected so far at each step. However, since the arms are adaptively selected--thereby yielding no n-iid data--conducting inference accurately is not straightforward. In particula r, sample averaging, which is used in the family of UCB and Thompson sampling (T S) algorithms, does not provide a good choice as it suffers from bias and a lack of good statistical properties (e.g. asymptotic normality). Our thesis in this paper is that more sophisticated inference schemes that take into account the a daptive nature of the sequentially collected data can unlock further performance gains, even though both UCB and TS type algorithms are optimal in the worst cas e. In particular, we propose a variant of TS-style algorithms--which we call dou bly adaptive TS--that leverages recent advances in causal inference and adaptive ly reweights the terms of a doubly robust estimator on the true mean reward of e ach arm. Through 20 synthetic domain experiments and a semi-synthetic experiment based on data from an A/B test of a web service, we demonstrate that using an a daptive inferential scheme (while still retaining the exploration efficacy of TS) provides clear benefits in online decision making: the proposed DATS algorithm

has superior empirical performance to existing baselines (UCB and TS) in terms of regret and sample complexity in identifying the best arm. In addition, we also provide a finite-time regret bound of doubly adaptive TS that matches (up to log factors) those of UCB and TS algorithms, thereby establishing that its improved practical benefits do not come at the expense of worst-case suboptimality.
****
Efficient Truncated Linear Regression with Unknown Noise Variance
Constantinos Daskalakis, Patroklos Stefanou, Rui Yao, Emmanouil Zampetakis
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Breaking the Dilemma of Medical Image-to-image Translation
Lingke Kong, Chenyu Lian, Detian Huang, zhenjiang li, Yanle Hu, Qichao Zhou
Supervised Pix2Pix and unsupervised Cycle-consistency are two modes that dominate the field of medical image-to-image translation. However, neither modes are ideal. The Pix2Pix mode has excellent performance. But it requires paired and well pixel-wise aligned images, which may not always be achievable due to respiratory motion or anatomy change between times that paired images are acquired. The Cycle-consistency mode is less stringent with training data and works well on unpaired or misaligned images. But its performance may not be optimal. In order to break the dilemma of the existing modes, we propose a new unsupervised mode called RegGAN for medical image-to-image translation. It is based on the theory of "loss-correction". In RegGAN, the misaligned target images are considered as noisy labels and the generator is trained with an additional registration network to fit the misaligned noise distribution adaptively. The goal is to search for the common optimal solution to both image-to-image translation and registration tasks. We incorporated RegGAN into a few state-of-the-art image-to-image translation methods and demonstrated that RegGAN could be easily combined with these methods to improve their performances. Such as a simple CycleGAN in our mode surpasses latest NICEGAN even though using less network parameters. Based on our results, RegGAN outperformed both Pix2Pix on aligned data and Cycle-consistency on misaligned or unpaired data. RegGAN is insensitive to noises which makes it a better choice for a wide range of scenarios, especially for medical image-to-image translation tasks in which well pixel-wise aligned data are not available. Code and dataset are available at https://github.com/Kid-Liet/Reg-GAN.
****
Temporally Abstract Partial Models
Khimya Khetarpal, Zafarali Ahmed, Gheorghe Comanici, Doina Precup
Humans and animals have the ability to reason and make predictions about different courses of action at many time scales. In reinforcement learning, option models (Sutton, Precup \& Singh, 1999; Precup, 2000) provide the framework for this kind of temporally abstract prediction and reasoning. Natural intelligent agents are also able to focus their attention on courses of action that are relevant or feasible in a given situation, sometimes termed affordable actions. In this paper, we define a notion of affordances for options, and develop temporally abstract partial option models, that take into account the fact that an option might be affordable only in certain situations. We analyze the trade-offs between estimation and approximation error in planning and learning when using such models, and identify some interesting special cases. Additionally, we empirically demonstrate the ability to learn both affordances and partial option models online resulting in improved sample efficiency and planning time in the Taxi domain.
****
TransMatcher: Deep Image Matching Through Transformers for Generalizable Person Re-identification
Shengcai Liao, Ling Shao
Transformers have recently gained increasing attention in computer vision. However, existing studies mostly use Transformers for feature representation learning, e.g. for image classification and dense predictions, and the generalizability

of Transformers is unknown. In this work, we further investigate the possibility of applying Transformers for image matching and metric learning given pairs of images. We find that the Vision Transformer (ViT) and the vanilla Transformer with decoders are not adequate for image matching due to their lack of image-to-image attention. Thus, we further design two naive solutions, i.e. query-gallery concatenation in ViT, and query-gallery cross-attention in the vanilla Transformer. The latter improves the performance, but it is still limited. This implies that the attention mechanism in Transformers is primarily designed for global feature aggregation, which is not naturally suitable for image matching. Accordingly, we propose a new simplified decoder, which drops the full attention implementation with the softmax weighting, keeping only the query-key similarity computation. Additionally, global max pooling and a multilayer perceptron (MLP) head are applied to decode the matching result. This way, the simplified decoder is computationally more efficient, while at the same time more effective for image matching. The proposed method, called TransMatcher, achieves state-of-the-art performance in generalizable person re-identification, with up to 6.1% and 5.7% performance gains in Rank-1 and mAP, respectively, on several popular datasets. Code is available at https://github.com/ShengcaiLiao/QAConv.
****
Multi-Objective SPIBB: Seldonian Offline Policy Improvement with Safety Constraints in Finite MDPs
harsh satija, Philip S. Thomas, Joelle Pineau, Romain Laroche
We study the problem of Safe Policy Improvement (SPI) under constraints in the offline Reinforcement Learning (RL) setting. We consider the scenario where: (i) we have a dataset collected under a known baseline policy, (ii) multiple reward signals are received from the environment inducing as many objectives to optimize. We present an SPI formulation for this RL setting that takes into account the preferences of the algorithm’s user for handling the trade-offs for different reward signals while ensuring that the new policy performs at least as well as the baseline policy along each individual objective. We build on traditional SPI algorithms and propose a novel method based on Safe Policy Iteration with Baseline Bootstrapping (SPIBB, Laroche et al., 2019) that provides high probability guarantees on the performance of the agent in the true environment. We show the effectiveness of our method on a synthetic grid-world safety task as well as in a real-world critical care context to learn a policy for the administration of IV fluids and vasopressors to treat sepsis.
****
Is Automated Topic Model Evaluation Broken? The Incoherence of Coherence
Alexander Hoyle, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan Boyd-Graber, Philip Resnik
Topic model evaluation, like evaluation of other unsupervised methods, can be contentious. However, the field has coalesced around automated estimates of topic coherence, which rely on the frequency of word co-occurrences in a reference corpus. Contemporary neural topic models surpass classical ones according to these metrics. At the same time, topic model evaluation suffers from a validation gap: automated coherence, developed for classical models, has not been validated using human experimentation for neural models. In addition, a meta-analysis of topic modeling literature reveals a substantial standardization gap in automated topic modeling benchmarks. To address the validation gap, we compare automated coherence with the two most widely accepted human judgment tasks: topic rating and word intrusion. To address the standardization gap, we systematically evaluate a dominant classical model and two state-of-the-art neural models on two commonly used datasets. Automated evaluations declare a winning model when corresponding human evaluations do not, calling into question the validity of fully automatic evaluations independent of human judgments.
****
INDIGO: GNN-Based Inductive Knowledge Graph Completion Using Pair-Wise Encoding
Shuwen Liu, Bernardo Grau, Ian Horrocks, Egor Kostylev
The aim of knowledge graph (KG) completion is to extend an incomplete KG with missing triples. Popular approaches based on graph embeddings typically work by fi

rst representing the KG in a vector space, and then applying a predefined scorin
g function to the resulting vectors to complete the KG. These approaches work we
ll in transductive settings, where predicted triples involve only constants seen
 during training; however, they are not applicable in inductive settings, where
the KG on which the model was trained is extended with new constants or merged w
ith other KGs. The use of Graph Neural Networks (GNNs) has recently been propose
d as a way to overcome these limitations; however, existing approaches do not fu
lly exploit the capabilities of GNNs and still rely on heuristics and ad-hoc sco
ring functions. In this paper, we propose a novel approach, where the KG is full
y encoded into a GNN in a transparent way, and where the predicted triples can b
e read out directly from the last layer of the GNN without the need for addition
al components or scoring functions. Our experiments show that our model outperfo
rms state-of-the-art approaches on inductive KG completion benchmarks.
****
Do Input Gradients Highlight Discriminative Features?
Harshay Shah, Prateek Jain, Praneeth Netrapalli
Post-hoc gradient-based interpretability methods [Simonyan et al., 2013, Smilkov
 et al., 2017] that provide instance-specific explanations of model predictions
are often based on assumption (A): magnitude of input gradients—gradients of log
its with respect to input—noisily highlight discriminative task-relevant feature
s. In this work, we test the validity of assumption (A) using a three-pronged ap
proach:1. We develop an evaluation framework, DiffROAR, to test assumption (A) o
n four image classification benchmarks. Our results suggest that (i) input gradi
ents of standard models (i.e., trained on original data) may grossly violate (A)
, whereas (ii) input gradients of adversarially robust models satisfy (A).2. We
then introduce BlockMNIST, an MNIST-based semi-real dataset, that by design enco
des a priori knowledge of discriminative features. Our analysis on BlockMNIST le
verages this information to validate as well as characterize differences between
 input gradient attributions of standard and robust models.3. Finally, we theore
tically prove that our empirical findings hold on a simplified version of the Bl
ockMNIST dataset. Specifically, we prove that input gradients of standard one-hi
dden-layer MLPs trained on this dataset do not highlight instance-specific signa
l coordinates, thus grossly violating assumption (A).Our findings motivate the n
eed to formalize and test common assumptions in interpretability in a falsifiabl
e manner [Leavitt and Morcos, 2020]. We believe that the DiffROAR evaluation fra
mework and BlockMNIST-based datasets can serve as sanity checks to audit instanc
e-specific interpretability methods; code and data available at https://github.c
om/harshays/inputgradients.
****
Improving Conditional Coverage via Orthogonal Quantile Regression
Shai Feldman, Stephen Bates, Yaniv Romano
We develop a method to generate prediction intervals that have a user-specified
coverage level across all regions of feature-space, a property called conditiona
l coverage. A typical approach to this task is to estimate the conditional quant
iles with quantile regression---it is well-known that this leads to correct cove
rage in the large-sample limit, although it may not be accurate in finite sample
s. We find in experiments that traditional quantile regression can have poor con
ditional coverage. To remedy this, we modify the loss function to promote indepe
ndence between the size of the intervals and the indicator of a miscoverage even
t. For the true conditional quantiles, these two quantities are independent (ort
hogonal), so the modified loss function continues to be valid. Moreover, we empi
rically show that the modified loss function leads to improved conditional cover
age, as evaluated by several metrics. We also introduce two new metrics that che
ck conditional coverage by looking at the strength of the dependence between the
 interval size and the indicator of miscoverage.
****
Minimizing Polarization and Disagreement in Social Networks via Link Recommendat
ion
Liwang Zhu, Qi Bao, Zhongzhi 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.
****

Adversarial Attacks on Black Box Video Classifiers: Leveraging the Power of Geometric Transformations
Shasha Li, Abhishek Aich, Shitong Zhu, Salman Asif, Chengyu Song, Amit Roy-Chowdhury, Srikanth Krishnamurthy
When compared to the image classification models, black-box adversarial attacks against video classification models have been largely understudied. This could be possible because, with video, the temporal dimension poses significant additional challenges in gradient estimation. Query-efficient black-box attacks rely on effectively estimated gradients towards maximizing the probability of misclassifying the target video. In this work, we demonstrate that such effective gradients can be searched for by parameterizing the temporal structure of the search space with geometric transformations. Specifically, we design a novel iterative algorithm GEOmetric TRAnsformed Perturbations (GEO-TRAP), for attacking video classification models. GEO-TRAP employs standard geometric transformation operations to reduce the search space for effective gradients into searching for a small group of parameters that define these operations. This group of parameters describes the geometric progression of gradients, resulting in a reduced and structured search space. Our algorithm inherently leads to successful perturbations with surprisingly few queries. For example, adversarial examples generated from GEO-TRAP have better attack success rates with ~73.55% fewer queries compared to the state-of-the-art method for video adversarial attacks on the widely used Jester dataset. Overall, our algorithm exposes vulnerabilities of diverse video classification models and achieves new state-of-the-art results under black-box settings on two large datasets.
****

Optimal Rates for Random Order Online Optimization
Uri Sherman, Tomer Koren, Yishay Mansour
We study online convex optimization in the random order model, recently proposed by Garber et al. (2020), where the loss functions may be chosen by an adversary, but are then presented to the online algorithm in a uniformly random order. Focusing on the scenario where the cumulative loss function is (strongly) convex, yet individual loss functions are smooth but might be non-convex, we give algorithms that achieve the optimal bounds and significantly outperform the results of Garber et al. (2020), completely removing the dimension dependence and improve their scaling with respect to the strong convexity parameter. Our analysis relies on novel connections between algorithmic stability and generalization for sampling without-replacement analogous to those studied in the with-replacement i.i.d. setting, as well as on a refined average stability analysis of stochastic gradient descent.
****

Discrete-Valued Neural Communication
Dianbo Liu, Alex M. Lamb, Kenji Kawaguchi, Anirudh Goyal ALIAS PARTH GOYAL, Chen Sun, Michael C. Mozer, Yoshua Bengio
Deep learning has advanced from fully connected architectures to structured models organized into components, e.g., the transformer composed of positional elements, modular architectures divided into slots, and graph neural nets made up of nodes. The nature of structured models is that communication among the components has a bottleneck, typically achieved by restricted connectivity and attention. In this work, we further tighten the bottleneck via discreteness of the representations transmitted between components. We hypothesize that this constraint serves as a useful form of inductive bias. Our hypothesis is motivated by past empirical work showing the benefits of discretization in non-structured architectures as well as our own theoretical results showing that discretization increases noise robustness and reduces the underlying dimensionality of the model. Building on an existing technique for discretization from the VQ-VAE, we consider multi-headed discretization with shared codebooks as the output of each architectural

component. One motivating intuition is human language in which communication occurs through multiple discrete symbols. This form of communication is hypothesized to facilitate transmission of information between functional components of the brain by providing a common interlingua, just as it does for human-to-human communication. Our experiments show that discrete-valued neural communication (DVNC) substantially improves systematic generalization in a variety of architectures—transformers, modular architectures, and graph neural networks. We also show that the DVNC is robust to the choice of hyperparameters, making the method useful in practice.
****

Skyformer: Remodel Self-Attention with Gaussian Kernel and Nystr\"om Method
Yifan Chen, Qi Zeng, Heng Ji, Yun Yang
Transformers are expensive to train due to the quadratic time and space complexity in the self-attention mechanism. On the other hand, although kernel machines suffer from the same computation bottleneck in pairwise dot products, several approximation schemes have been successfully incorporated to considerably reduce their computational cost without sacrificing too much accuracy. In this work, we leverage the computation methods for kernel machines to alleviate the high computational cost and introduce Skyformer, which replaces the softmax structure with a Gaussian kernel to stabilize the model training and adapts the Nyström method to a non-positive semidefinite matrix to accelerate the computation. We further conduct theoretical analysis by showing that the matrix approximation error of our proposed method is small in the spectral norm. Experiments on Long Range Arena benchmark show that the proposed method is sufficient in getting comparable or even better performance than the full self-attention while requiring fewer computation resources.
****

TransMIL: Transformer based Correlated Multiple Instance Learning for Whole Slide Image Classification
Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, yongbing zhang
Multiple instance learning (MIL) is a powerful tool to solve the weakly supervised classification in whole slide image (WSI) based pathology diagnosis. However, the current MIL methods are usually based on independent and identical distribution hypothesis, thus neglect the correlation among different instances. To address this problem, we proposed a new framework, called correlated MIL, and provided a proof for convergence. Based on this framework, we devised a Transformer based MIL (TransMIL), which explored both morphological and spatial information. The proposed TransMIL can effectively deal with unbalanced/balanced and binary/multiple classification with great visualization and interpretability. We conducted various experiments for three different computational pathology problems and achieved better performance and faster convergence compared with state-of-the-art methods. The test AUC for the binary tumor classification can be up to 93.09% over CAMELYON16 dataset. And the AUC over the cancer subtypes classification can be up to 96.03% and 98.82% over TCGA-NSCLC dataset and TCGA-RCC dataset, respectively. Implementation is available at: https://github.com/szc19990412/TransMIL.
****

Multi-view Contrastive Graph Clustering
ErLin Pan, Zhao Kang
With the explosive growth of information technology, multi-view graph data have become increasingly prevalent and valuable. Most existing multi-view clustering techniques either focus on the scenario of multiple graphs or multi-view attributes. In this paper, we propose a generic framework to cluster multi-view attributed graph data. Specifically, inspired by the success of contrastive learning, we propose multi-view contrastive graph clustering (MCGC) method to learn a consensus graph since the original graph could be noisy or incomplete and is not directly applicable. Our method composes of two key steps: we first filter out the undesirable high-frequency noise while preserving the graph geometric features via graph filtering and obtain a smooth representation of nodes; we then learn a consensus graph regularized by graph contrastive loss. Results on several benchma

rk datasets show the superiority of our method with respect to state-of-the-art approaches. In particular, our simple approach outperforms existing deep learnin g-based methods.
****
Inverse-Weighted Survival Games
Xintian Han, Mark Goldstein, Aahlad Puli, Thomas Wies, Adler Perotte, Rajesh Ran ganath
Deep models trained through maximum likelihood have achieved state-of-the-art re sults for survival analysis. Despite this training scheme, practitioners evaluat e models under other criteria, such as binary classification losses at a chosen set of time horizons, e.g. Brier score (BS) and Bernoulli log likelihood (BLL). Models trained with maximum likelihood may have poor BS or BLL since maximum lik elihood does not directly optimize these criteria. Directly optimizing criteria like BS requires inverse-weighting by the censoring distribution. However, estim ating the censoring model under these metrics requires inverse-weighting by the failure distribution. The objective for each model requires the other, but neith er are known. To resolve this dilemma, we introduce Inverse-Weighted Survival Ga mes. In these games, objectives for each model are built from re-weighted estima tes featuring the other model, where the latter is held fixed during training. W hen the loss is proper, we show that the games always have the true failure and censoring distributions as a stationary point. This means models in the game do not leave the correct distributions once reached. We construct one case where th is stationary point is unique. We show that these games optimize BS on simulatio ns and then apply these principles on real world cancer and critically-ill patie nt data.
****
Generalization Bounds for Meta-Learning via PAC-Bayes and Uniform Stability
Alec Farid, Anirudha Majumdar
We are motivated by the problem of providing strong generalization guarantees in the context of meta-learning. Existing generalization bounds are either challen ging to evaluate or provide vacuous guarantees in even relatively simple setting s. We derive a probably approximately correct (PAC) bound for gradient-based met a-learning using two different generalization frameworks in order to deal with t he qualitatively different challenges of generalization at the "base" and "meta" levels. We employ bounds for uniformly stable algorithms at the base level and bounds from the PAC-Bayes framework at the meta level. The result of this approa ch is a novel PAC bound that is tighter when the base learner adapts quickly, wh ich is precisely the goal of meta-learning. We show that our bound provides a ti ghter guarantee than other bounds on a toy non-convex problem on the unit sphere and a text-based classification example. We also present a practical regulariza tion scheme motivated by the bound in settings where the bound is loose and demo nstrate improved performance over baseline techniques.
****
Parallel Bayesian Optimization of Multiple Noisy Objectives with Expected Hyperv olume Improvement
Samuel Daulton, Maximilian Balandat, Eytan Bakshy
Optimizing multiple competing black-box objectives is a challenging problem in m any fields, including science, engineering, and machine learning. Multi-objectiv e Bayesian optimization (MOBO) is a sample-efficient approach for identifying th e optimal trade-offs between the objectives. However, many existing methods perf orm poorly when the observations are corrupted by noise. We propose a novel acqu isition function, NEHVI, that overcomes this important practical limitation by a pplying a Bayesian treatment to the popular expected hypervolume improvement (EH VI) criterion and integrating over this uncertainty in the Pareto frontier. We a rgue that, even in the noiseless setting, generating multiple candidates in para llel is an incarnation of EHVI with uncertainty in the Pareto frontier and there fore can be addressed using the same underlying technique. Through this lens, we derive a natural parallel variant, qNEHVI, that reduces computational complexit y of parallel EHVI from exponential to polynomial with respect to the batch size . qNEHVI is one-step Bayes-optimal for hypervolume maximization in both noisy an

d noiseless environments, and we show that it can be optimized effectively with gradient-based methods via sample average approximation. Empirically, we demonst rate not only that qNEHVI is substantially more robust to observation noise than existing MOBO approaches, but also that it achieves state-of-the-art optimizati on performance and competitive wall-times in large-batch environments.
****
Evolution Gym: A Large-Scale Benchmark for Evolving Soft Robots
Jagdeep Bhatia, Holly Jackson, Yunsheng Tian, Jie Xu, Wojciech Matusik
Both the design and control of a robot play equally important roles in its task performance. However, while optimal control is well studied in the machine learn ing and robotics community, less attention is placed on finding the optimal robo t design. This is mainly because co-optimizing design and control in robotics is characterized as a challenging problem, and more importantly, a comprehensive e valuation benchmark for co-optimization does not exist. In this paper, we propos e Evolution Gym, the first large-scale benchmark for co-optimizing the design an d control of soft robots. In our benchmark, each robot is composed of different types of voxels (e.g., soft, rigid, actuators), resulting in a modular and expre ssive robot design space. Our benchmark environments span a wide range of tasks, including locomotion on various types of terrains and manipulation. Furthermore , we develop several robot co-evolution algorithms by combining state-of-the-art design optimization methods and deep reinforcement learning techniques. Evaluat ing the algorithms on our benchmark platform, we observe robots exhibiting incre asingly complex behaviors as evolution progresses, with the best evolved designs solving many of our proposed tasks. Additionally, even though robot designs are evolved autonomously from scratch without prior knowledge, they often grow to r esemble existing natural creatures while outperforming hand-designed robots. Nev ertheless, all tested algorithms fail to find robots that succeed in our hardest environments. This suggests that more advanced algorithms are required to explo re the high-dimensional design space and evolve increasingly intelligent robots -- an area of research in which we hope Evolution Gym will accelerate progress. Our website with code, environments, documentation, and tutorials is available a t http://evogym.csail.mit.edu/.
****
On Calibration and Out-of-Domain Generalization
Yoav Wald, Amir Feder, Daniel Greenfeld, Uri Shalit
Out-of-domain (OOD) generalization is a significant challenge for machine learni ng models. Many techniques have been proposed to overcome this challenge, often focused on learning models with certain invariance properties. In this work, we draw a link between OOD performance and model calibration, arguing that calibrat ion across multiple domains can be viewed as a special case of an invariant repr esentation leading to better OOD generalization. Specifically, we show that unde r certain conditions, models which achieve \emph{multi-domain calibration} are p rovably free of spurious correlations. This leads us to propose multi-domain cal ibration as a measurable and trainable surrogate for the OOD performance of a cl assifier. We therefore introduce methods that are easy to apply and allow practi tioners to improve multi-domain calibration by training or modifying an existing model, leading to better performance on unseen domains. Using four datasets fro m the recently proposed WILDS OOD benchmark, as well as the Colored MNIST, we de monstrate that training or tuning models so they are calibrated across multiple domains leads to significantly improved performance on unseen test domains. We b elieve this intriguing connection between calibration and OOD generalization is promising from both a practical and theoretical point of view.
****
On the Convergence and Sample Efficiency of Variance-Reduced Policy Gradient Met hod
Junyu Zhang, Chengzhuo Ni, zheng Yu, Csaba Szepesvari, Mengdi 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.

****
Circa: Stochastic ReLUs for Private Deep Learning
Zahra Ghodsi, Nandan Kumar Jha, Brandon Reagen, Siddharth Garg
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Reinforcement Learning in Reward-Mixing MDPs
Jeongyeol Kwon, Yonathan Efroni, Constantine Caramanis, 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-authors prior to requesting a name change in the electronic proceedings.
****
A Gang of Adversarial Bandits
Mark Herbster, Stephen Pasteris, Fabio Vitale, Massimiliano Pontil
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Explaining Hyperparameter Optimization via Partial Dependence Plots
Julia Moosbauer, Julia Herbinger, Giuseppe Casalicchio, Marius Lindauer, Bernd Bischl
Automated hyperparameter optimization (HPO) can support practitioners to obtain peak performance in machine learning models.However, there is often a lack of valuable insights into the effects of different hyperparameters on the final model performance.This lack of explainability makes it difficult to trust and understand the automated HPO process and its results.We suggest using interpretable machine learning (IML) to gain insights from the experimental data obtained during HPO with Bayesian optimization (BO).BO tends to focus on promising regions with potential high-performance configurations and thus induces a sampling bias.Hence, many IML techniques, such as the partial dependence plot (PDP), carry the risk of generating biased interpretations.By leveraging the posterior uncertainty of the BO surrogate model, we introduce a variant of the PDP with estimated confidence bands.We propose to partition the hyperparameter space to obtain more confident and reliable PDPs in relevant sub-regions.In an experimental study, we provide quantitative evidence for the increased quality of the PDPs within sub-regions.
****
Robustifying Algorithms of Learning Latent Trees with Vector Variables
Fengzhuo Zhang, Vincent Tan
We consider learning the structures of Gaussian latent tree models with vector observations when a subset of them are arbitrarily corrupted. First, we present the sample complexities of Recursive Grouping (RG) and Chow-Liu Recursive Grouping (CLRG) without the assumption that the effective depth is bounded in the number of observed nodes, significantly generalizing the results in Choi et al. (2011). We show that Chow-Liu initialization in CLRG greatly reduces the sample complexity of RG from being exponential in the diameter of the tree to only logarithmic in the diameter for the hidden Markov model (HMM). Second, we robustify RG, CLRG, Neighbor Joining (NJ) and Spectral NJ (SNJ) by using the truncated inner product. These robustified algorithms can tolerate a number of corruptions up to the square root of the number of clean samples. Finally, we derive the first known instance-dependent impossibility result for structure learning of latent trees. The optimalities of the robust version of CLRG and NJ are verified by comparing their sample complexities and the impossibility result.
****
Representation Learning on Spatial Networks
Zheng Zhang, Liang Zhao

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Continuous-time edge modelling using non-parametric point processes
Xuhui Fan, Bin Li, Feng Zhou, Scott SIsson
The mutually-exciting Hawkes process (ME-HP) is a natural choice to model reciprocity, which is an important attribute of continuous-time edge (dyadic) data. However, existing ways of implementing the ME-HP for such data are either inflexible, as the exogenous (background) rate functions are typically constant and the endogenous (excitation) rate functions are specified parametrically, or inefficient, as inference usually relies on Markov chain Monte Carlo methods with high computational costs. To address these limitations, we discuss various approaches to model design, and develop three variants of non-parametric point processes for continuous-time edge modelling (CTEM). The resulting models are highly adaptable as they generate intensity functions through sigmoidal Gaussian processes, and so provide greater modelling flexibility than parametric forms. The models are implemented via a fast variational inference method enabled by a novel edge modelling construction. The superior performance of the proposed CTEM models is demonstrated through extensive experimental evaluations on four real-world continuous-time edge data sets.
****
Deep inference of latent dynamics with spatio-temporal super-resolution using selective backpropagation through time
Feng Zhu, Andrew Sedler, Harrison A Grier, Nauman Ahad, Mark Davenport, Matthew Kaufman, Andrea Giovannucci, Chethan Pandarinath
Modern neural interfaces allow access to the activity of up to a million neurons within brain circuits. However, bandwidth limits often create a trade-off between greater spatial sampling (more channels or pixels) and the temporal frequency of sampling. Here we demonstrate that it is possible to obtain spatio-temporal super-resolution in neuronal time series by exploiting relationships among neurons, embedded in latent low-dimensional population dynamics. Our novel neural network training strategy, selective backpropagation through time (SBTT), enables learning of deep generative models of latent dynamics from data in which the set of observed variables changes at each time step. The resulting models are able to infer activity for missing samples by combining observations with learned latent dynamics. We test SBTT applied to sequential autoencoders and demonstrate more efficient and higher-fidelity characterization of neural population dynamics in electrophysiological and calcium imaging data. In electrophysiology, SBTT enables accurate inference of neuronal population dynamics with lower interface bandwidths, providing an avenue to significant power savings for implanted neuroelectronic interfaces. In applications to two-photon calcium imaging, SBTT accurately uncovers high-frequency temporal structure underlying neural population activity, substantially outperforming the current state-of-the-art. Finally, we demonstrate that performance could be further improved by using limited, high-bandwidth sampling to pretrain dynamics models, and then using SBTT to adapt these models for sparsely-sampled data.
****
Memory-efficient Patch-based Inference for Tiny Deep Learning
Ji Lin, Wei-Ming Chen, Han Cai, Chuang Gan, Song Han
Tiny deep learning on microcontroller units (MCUs) is challenging due to the limited memory size. We find that the memory bottleneck is due to the imbalanced memory distribution in convolutional neural network (CNN) designs: the first several blocks have an order of magnitude larger memory usage than the rest of the network. To alleviate this issue, we propose a generic patch-by-patch inference scheduling, which operates only on a small spatial region of the feature map and significantly cuts down the peak memory. However, naive implementation brings overlapping patches and computation overhead. We further propose receptive field redistribution to shift the receptive field and FLOPs to the later stage and redu

ce the computation overhead. Manually redistributing the receptive field is diff icult. We automate the process with neural architecture search to jointly optimi ze the neural architecture and inference scheduling, leading to MCUNetV2. Patch-based inference effectively reduces the peak memory usage of existing networks b y4-8×. Co-designed with neural networks, MCUNetV2 sets a record ImageNetaccurac y on MCU (71.8%) and achieves >90% accuracy on the visual wake words dataset und er only 32kB SRAM. MCUNetV2 also unblocks object detection on tiny devices, achi eving 16.9% higher mAP on Pascal VOC compared to the state-of-the-art result. Ou r study largely addressed the memory bottleneck in tinyML and paved the way for various vision applications beyond image classification.
****
Self-Interpretable Model with Transformation Equivariant Interpretation
Yipei Wang, Xiaoqian Wang
With the proliferation of machine learning applications in the real world, the d emand for explaining machine learning predictions continues to grow especially i n high-stakes fields. Recent studies have found that interpretation methods can be sensitive and unreliable, where the interpretations can be disturbed by pertu rbations or transformations of input data. To address this issue, we propose to learn robust interpretation through transformation equivariant regularization in a self-interpretable model. The resulting model is capable of capturing valid i nterpretation that is equivariant to geometric transformations. Moreover, since our model is self-interpretable, it enables faithful interpretations that reflec t the true predictive mechanism. Unlike existing self-interpretable models, whic h usually sacrifice expressive power for the sake of interpretation quality, our model preserves the high expressive capability comparable to the state-of-the-a rt deep learning models in complex tasks, while providing visualizable and faith ful high-quality interpretation. We compare with various related methods and val idate the interpretation quality and consistency of our model.
****
Solving Min-Max Optimization with Hidden Structure via Gradient Descent Ascent
Emmanouil-Vasileios Vlatakis-Gkaragkounis, Lampros Flokas, Georgios Piliouras
Many recent AI architectures are inspired by zero-sum games, however, the behavi or of their dynamics is still not well understood. Inspired by this, we study st andard gradient descent ascent (GDA) dynamics in a specific class of non-convex non-concave zero-sum games, that we call hidden zero-sum games. In this class, p layers control the inputs of smooth but possibly non-linear functions whose outp uts are being applied as inputs to a convex-concave game. Unlike general zero-su m games, these games have a well-defined notion of solution; outcomes that imple ment the von-Neumann equilibrium of the ``hidden" convex-concave game. We provid e conditions under which vanilla GDA provably converges not merely to local Nash , but the actual von-Neumann solution. If the hidden game lacks strict convexity properties, GDA may fail to converge to any equilibrium, however, by applying s tandard regularization techniques we can prove convergence to a von-Neumann solu tion of a slightly perturbed zero-sum game. Our convergence results are non-loca l despite working in the setting of non-convex non-concave games. Critically, un der proper assumptions we combine the Center-Stable Manifold Theorem along with novel type of initialization dependent Lyapunov functions to prove that almost a ll initial conditions converge to the solution. Finally, we discuss diverse appl ications of our framework ranging from generative adversarial networks to evolut ionary biology.
****
Preserved central model for faster bidirectional compression in distributed sett ings
Constantin Philippenko, Aymeric Dieuleveut
We develop a new approach to tackle communication constraints in a distributed l earning problem with a central server. We propose and analyze a new algorithm th at performs bidirectional compression and achieves the same convergence rate as algorithms using only uplink (from the local workers to the central server) comp ression. To obtain this improvement, we design MCM, an algorithm such that the d ownlink compression only impacts local models, while the global model is preserv

ed. As a result, and contrary to previous works, the gradients on local servers are computed on perturbed models. Consequently, convergence proofs are more challenging and require a precise control of this perturbation. To ensure it, MCM additionally combines model compression with a memory mechanism. This analysis opens new doors, e.g. incorporating worker dependent randomized-models and partial participation.
****
Understanding Instance-based Interpretability of Variational Auto-Encoders
Zhifeng Kong, Kamalika Chaudhuri
Instance-based interpretation methods have been widely studied for supervised learning methods as they help explain how black box neural networks predict. However, instance-based interpretations remain ill-understood in the context of unsupervised learning. In this paper, we investigate influence functions [Koh and Liang, 2017], a popular instance-based interpretation method, for a class of deep generative models called variational auto-encoders (VAE). We formally frame the counter-factual question answered by influence functions in this setting, and through theoretical analysis, examine what they reveal about the impact of training samples on classical unsupervised learning methods. We then introduce VAE- TracIn, a computationally efficient and theoretically sound solution based on Pruthi et al. [2020], for VAEs. Finally, we evaluate VAE-TracIn on several real world datasets with extensive quantitative and qualitative analysis.
****
Voxel-based 3D Detection and Reconstruction of Multiple Objects from a Single Image
Feng Liu, Xiaoming Liu
Inferring 3D locations and shapes of multiple objects from a single 2D image is a long-standing objective of computer vision. Most of the existing works either predict one of these 3D properties or focus on solving both for a single object. One fundamental challenge lies in how to learn an effective representation of the image that is well-suited for 3D detection and reconstruction. In this work, we propose to learn a regular grid of 3D voxel features from the input image which is aligned with 3D scene space via a 3D feature lifting operator. Based on the 3D voxel features, our novel CenterNet-3D detection head formulates the 3D detection as keypoint detection in the 3D space. Moreover, we devise an efficient coarse-to-fine reconstruction module, including coarse-level voxelization and a novel local PCA-SDF shape representation, which enables fine detail reconstruction and two orders of magnitude faster inference than prior methods. With complementary supervision from both 3D detection and reconstruction, one enables the 3D voxel features to be geometry and context preserving, benefiting both tasks. The effectiveness of our approach is demonstrated through 3D detection and reconstruction on single-object and multiple-object scenarios.
****
Test-Time Classifier Adjustment Module for Model-Agnostic Domain Generalization
Yusuke Iwasawa, Yutaka Matsuo
This paper presents a new algorithm for domain generalization (DG), \textit{test-time template adjuster (T3A)}, aiming to robustify a model to unknown distribution shift. Unlike existing methods that focus on \textit{training phase}, our method focuses \textit{test phase}, i.e., correcting its prediction by itself during test time. Specifically, T3A adjusts a trained linear classifier (the last layer of deep neural networks) with the following procedure: (1) compute a pseudo-prototype representation for each class using online unlabeled data augmented by the base classifier trained in the source domains, (2) and then classify each sample based on its distance to the pseudo-prototypes. T3A is back-propagation-free and modifies only the linear layer; therefore, the increase in computational cost during inference is negligible and avoids the catastrophic failure might caused by stochastic optimization. Despite its simplicity, T3A can leverage knowledge about the target domain by using off-the-shelf test-time data and improve performance. We tested our method on four domain generalization benchmarks, namely PACS, VLCS, OfficeHome, and TerraIncognita, along with various backbone networks including ResNet18, ResNet50, Big Transfer (BiT), Vision Transformers (ViT),

and MLP-Mixer. The results show T3A stably improves performance on unseen domain
s across choices of backbone networks, and outperforms existing domain generaliz
ation methods.
****
Luna: Linear Unified Nested Attention
Xuezhe Ma, Xiang Kong, Sinong Wang, Chunting Zhou, Jonathan May, Hao Ma, Luke Ze
ttlemoyer
The quadratic computational and memory complexities of the Transformer's attenti
on mechanism have limited its scalability for modeling long sequences. In this
paper, we propose Luna, a linear unified nested attention mechanism that approxi
mates softmax attention with two nested linear attention functions, yielding onl
y linear (as opposed to quadratic) time and space complexity. Specifically, with
 the first attention function, Luna packs the input sequence into a sequence of
fixed length. Then, the packed sequence is unpacked using the second attention f
unction. As compared to a more traditional attention mechanism, Luna introduces
an additional sequence with a fixed length as input and an additional correspond
ing output, which allows Luna to perform attention operation linearly, while als
o storing adequate contextual information. We perform extensive evaluations on t
hree benchmarks of sequence modeling tasks: long-context sequence modelling, neu
ral machine translation and masked language modeling for large-scale pretraining
. Competitive or even better experimental results demonstrate both the effective
ness and efficiency of Luna compared to a variety of strong baseline methods inc
luding the full-rank attention and other efficient sparse and dense attention me
thods.
****
Iterative Causal Discovery in the Possible Presence of Latent Confounders and Se
lection Bias
Raanan Y. Rohekar, Shami Nisimov, Yaniv Gurwicz, Gal Novik
We present a sound and complete algorithm, called iterative causal discovery (IC
D), for recovering causal graphs in the presence of latent confounders and selec
tion bias. ICD relies on the causal Markov and faithfulness assumptions and reco
vers the equivalence class of the underlying causal graph. It starts with a comp
lete graph, and consists of a single iterative stage that gradually refines this
 graph by identifying conditional independence (CI) between connected nodes. Ind
ependence and causal relations entailed after any iteration are correct, renderi
ng ICD anytime. Essentially, we tie the size of the CI conditioning set to its d
istance on the graph from the tested nodes, and increase this value in the succe
ssive iteration. Thus, each iteration refines a graph that was recovered by prev
ious iterations having smaller conditioning sets---a higher statistical power---
which contributes to stability. We demonstrate empirically that ICD requires sig
nificantly fewer CI tests and learns more accurate causal graphs compared to FCI
, FCI+, and RFCI algorithms.
****
Hindsight Task Relabelling: Experience Replay for Sparse Reward Meta-RL
Charles Packer, Pieter Abbeel, Joseph E. Gonzalez
Meta-reinforcement learning (meta-RL) has proven to be a successful framework fo
r leveraging experience from prior tasks to rapidly learn new related tasks, how
ever, current meta-RL approaches struggle to learn in sparse reward environments
. Although existing meta-RL algorithms can learn strategies for adapting to new
sparse reward tasks, the actual adaptation strategies are learned using hand-sha
ped reward functions, or require simple environments where random exploration is
 sufficient to encounter sparse reward. In this paper we present a formulation o
f hindsight relabelling for meta-RL, which relabels experience during meta-train
ing to enable learning to learn entirely using sparse reward. We demonstrate the
 effectiveness of our approach on a suite of challenging sparse reward environme
nts that previously required dense reward during meta-training to solve. Our app
roach solves these environments using the true sparse reward function, with perf
ormance comparable to training with a proxy dense reward function.
****
A Bayesian-Symbolic Approach to Reasoning and Learning in Intuitive Physics

Kai Xu, Akash Srivastava, Dan Gutfreund, Felix Sosa, Tomer Ullman, Josh Tenenbaum, Charles Sutton

Humans can reason about intuitive physics in fully or partially observed environments even after being exposed to a very limited set of observations. This sample-efficient intuitive physical reasoning is considered a core domain of human common sense knowledge. One hypothesis to explain this remarkable capacity, posits that humans quickly learn approximations to the laws of physics that govern the dynamics of the environment. In this paper, we propose a Bayesian-symbolic framework (BSP) for physical reasoning and learning that is close to human-level sample-efficiency and accuracy. In BSP, the environment is represented by a top-down generative model of entities, which are assumed to interact with each other under unknown force laws over their latent and observed properties. BSP models each of these entities as random variables, and uses Bayesian inference to estimate their unknown properties. For learning the unknown forces, BSP leverages symbolic regression on a novel grammar of Newtonian physics in a bilevel optimization setup. These inference and regression steps are performed in an iterative manner using expectation-maximization, allowing BSP to simultaneously learn force laws while maintaining uncertainty over entity properties. We show that BSP is more sample-efficient compared to neural alternatives on controlled synthetic datasets, demonstrate BSP's applicability to real-world common sense scenes and study BSP's performance on tasks previously used to study human physical reasoning.

****

Associating Objects with Transformers for Video Object Segmentation

Zongxin Yang, Yunchao Wei, Yi Yang

This paper investigates how to realize better and more efficient embedding learning to tackle the semi-supervised video object segmentation under challenging multi-object scenarios. The state-of-the-art methods learn to decode features with a single positive object and thus have to match and segment each target separately under multi-object scenarios, consuming multiple times computing resources. To solve the problem, we propose an Associating Objects with Transformers (AOT) approach to match and decode multiple objects uniformly. In detail, AOT employs an identification mechanism to associate multiple targets into the same high-dimensional embedding space. Thus, we can simultaneously process multiple objects' matching and segmentation decoding as efficiently as processing a single object. For sufficiently modeling multi-object association, a Long Short-Term Transformer is designed for constructing hierarchical matching and propagation. We conduct extensive experiments on both multi-object and single-object benchmarks to examine AOT variant networks with different complexities. Particularly, our R50-AOT-L outperforms all the state-of-the-art competitors on three popular benchmarks, i.e., YouTube-VOS (84.1% J&F), DAVIS 2017 (84.9%), and DAVIS 2016 (91.1%), while keeping more than 3X faster multi-object run-time. Meanwhile, our AOT-T can maintain real-time multi-object speed on the above benchmarks. Based on AOT, we ranked 1st in the 3rd Large-scale VOS Challenge.

****

Automatic Symmetry Discovery with Lie Algebra Convolutional Network

Nima Dehmamy, Robin Walters, Yanchen Liu, Dashun Wang, Rose Yu

Existing equivariant neural networks require prior knowledge of the symmetry group and discretization for continuous groups. We propose to work with Lie algebras (infinitesimal generators) instead of Lie groups. Our model, the Lie algebra convolutional network (L-conv) can automatically discover symmetries and does not require discretization of the group. We show that L-conv can serve as a building block to construct any group equivariant feedforward architecture. Both CNNs and Graph Convolutional Networks can be expressed as L-conv with appropriate groups. We discover direct connections between L-conv and physics: (1) group invariant loss generalizes field theory (2) Euler-Lagrange equation measures the robustness, and (3) equivariance leads to conservation laws and Noether current. These connections open up new avenues for designing more general equivariant networks and applying them to important problems in physical sciences.

****

Zero Time Waste: Recycling Predictions in Early Exit Neural Networks

Maciej Wo■czyk, Bartosz Wójcik, Klaudia Ba■azy, Igor T Podolak, Jacek Tabor, Mar
ek ■mieja, Tomasz Trzcinski
The problem of reducing processing time of large deep learning models is a funda
mental challenge in many real-world applications. Early exit methods strive towa
rds this goal by attaching additional Internal Classifiers (ICs) to intermediate
 layers of a neural network. ICs can quickly return predictions for easy example
s and, as a result, reduce the average inference time of the whole model. Howeve
r, if a particular IC does not decide to return an answer early, its predictions
 are discarded, with its computations effectively being wasted. To solve this is
sue, we introduce Zero Time Waste (ZTW), a novel approach in which each IC reuse
s predictions returned by its predecessors by (1) adding direct connections betw
een ICs and (2) combining previous outputs in an ensemble-like manner. We conduc
t extensive experiments across various datasets and architectures to demonstrate
 that ZTW achieves a significantly better accuracy vs. inference time trade-off
than other recently proposed early exit methods.
****
On Model Calibration for Long-Tailed Object Detection and Instance Segmentation
Tai-Yu Pan, Cheng Zhang, Yandong Li, Hexiang Hu, Dong Xuan, Soravit Changpinyo,
Boqing Gong, Wei-Lun Chao
Vanilla models for object detection and instance segmentation suffer from the he
avy bias toward detecting frequent objects in the long-tailed setting. Existing
methods address this issue mostly during training, e.g., by re-sampling or re-we
ighting. In this paper, we investigate a largely overlooked approach --- post-pr
ocessing calibration of confidence scores. We propose NorCal, Normalized Calibra
tion for long-tailed object detection and instance segmentation, a simple and st
raightforward recipe that reweighs the predicted scores of each class by its tra
ining sample size. We show that separately handling the background class and nor
malizing the scores over classes for each proposal are keys to achieving superio
r performance. On the LVIS dataset, NorCal can effectively improve nearly all th
e baseline models not only on rare classes but also on common and frequent class
es. Finally, we conduct extensive analysis and ablation studies to offer insigh
ts into various modeling choices and mechanisms of our approach. Our code is pub
licly available at https://github.com/tydpan/NorCal.
****
ReSSL: Relational Self-Supervised Learning with Weak Augmentation
Mingkai Zheng, Shan You, Fei Wang, Chen Qian, Changshui Zhang, Xiaogang Wang, Ch
ang Xu
Self-supervised Learning (SSL) including the mainstream contrastive learning has
 achieved great success in learning visual representations without data annotati
ons. However, most of methods mainly focus on the instance level information (\i
e, the different augmented images of the same instance should have the same feat
ure or cluster into the same class), but there is a lack of attention on the rel
ationships between different instances. In this paper, we introduced a novel SSL
 paradigm, which we term as relational self-supervised learning (ReSSL) framewo
rk that learns representations by modeling the relationship between different in
stances. Specifically, our proposed method employs sharpened distribution of pai
rwise similarities among different instances as \textit{relation} metric, which
is thus utilized to match the feature embeddings of different augmentations. Mor
eover, to boost the performance, we argue that weak augmentations matter to repr
esent a more reliable relation, and leverage momentum strategy for practical eff
iciency. Experimental results show that our proposed ReSSL significantly outperf
orms the previous state-of-the-art algorithms in terms of both performance and t
raining efficiency.
****
Learning to See by Looking at Noise
Manel Baradad Jurjo, Jonas Wulff, Tongzhou Wang, Phillip Isola, Antonio Torralba
Current vision systems are trained on huge datasets, and these datasets come wit
h costs: curation is expensive, they inherit human biases, and there are concern
s over privacy and usage rights. To counter these costs, interest has surged in
learning from cheaper data sources, such as unlabeled images. In this paper we g

o a step further and ask if we can do away with real image datasets entirely, instead learning from procedural noise processes. We investigate a suite of image generation models that produce images from simple random processes. These are then used as training data for a visual representation learner with a contrastive loss. In particular, we study statistical image models, randomly initialized deep generative models, and procedural graphics models.Our findings show that it is important for the noise to capture certain structural properties of real data but that good performance can be achieved even with processes that are far from realistic. We also find that diversity is a key property to learn good representations.
****
Explicit loss asymptotics in the gradient descent training of neural networks
Maksim Velikanov, Dmitry Yarotsky
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Test-Time Personalization with a Transformer for Human Pose Estimation
Yizhuo Li, Miao Hao, Zonglin Di, Nitesh Bharadwaj Gundavarapu, Xiaolong Wang
We propose to personalize a 2D human pose estimator given a set of test images of a person without using any manual annotations. While there is a significant advancement in human pose estimation, it is still very challenging for a model to generalize to different unknown environments and unseen persons. Instead of using a fixed model for every test case, we adapt our pose estimator during test time to exploit person-specific information. We first train our model on diverse data with both a supervised and a self-supervised pose estimation objectives jointly. We use a Transformer model to build a transformation between the self-supervised keypoints and the supervised keypoints. During test time, we personalize and adapt our model by fine-tuning with the self-supervised objective. The pose is then improved by transforming the updated self-supervised keypoints. We experiment with multiple datasets and show significant improvements on pose estimations with our self-supervised personalization. Project page with code is available at https://liyz15.github.io/TTP/.
****
Towards Scalable Unpaired Virtual Try-On via Patch-Routed Spatially-Adaptive GAN
Zhenyu Xie, Zaiyu Huang, Fuwei Zhao, Haoye Dong, Michael Kampffmeyer, Xiaodan Liang
Image-based virtual try-on is one of the most promising applications of human-centric image generation due to its tremendous real-world potential. Yet, as most try-on approaches fit in-shop garments onto a target person, they require the laborious and restrictive construction of a paired training dataset, severely limiting their scalability. While a few recent works attempt to transfer garments directly from one person to another, alleviating the need to collect paired datasets, their performance is impacted by the lack of paired (supervised) information. In particular, disentangling style and spatial information of the garment becomes a challenge, which existing methods either address by requiring auxiliary data or extensive online optimization procedures, thereby still inhibiting their scalability. To achieve a scalable virtual try-on system that can transfer arbitrary garments between a source and a target person in an unsupervised manner, we thus propose a texture-preserving end-to-end network, the PAtch-routed SpaTially-Adaptive GAN (PASTA-GAN), that facilitates real-world unpaired virtual try-on. Specifically, to disentangle the style and spatial information of each garment, PASTA-GAN consists of an innovative patch-routed disentanglement module for successfully retaining garment texture and shape characteristics. Guided by the source person's keypoints, the patch-routed disentanglement module first decouples garments into normalized patches, thus eliminating the inherent spatial information of the garment, and then reconstructs the normalized patches to the warped garment complying with the target person pose. Given the warped garment, PASTA-GAN further introduces novel spatially-adaptive residual blocks that guide the ge

nerator to synthesize more realistic garment details. Extensive comparisons with paired and unpaired approaches demonstrate the superiority of PASTA-GAN, highlighting its ability to generate high-quality try-on images when faced with a large variety of garments(e.g. vests, shirts, pants), taking a crucial step towards real-world scalable try-on.
****
Bias Out-of-the-Box: An Empirical Analysis of Intersectional Occupational Biases in Popular Generative Language Models
Hannah Rose Kirk, Yennie Jun, Filippo Volpin, Haider Iqbal, Elias Benussi, Frederic Dreyer, Aleksandar Shtedritski, Yuki Asano
The capabilities of natural language models trained on large-scale data have increased immensely over the past few years. Open source libraries such as HuggingFace have made these models easily available and accessible. While prior research has identified biases in large language models, this paper considers biases contained in the most popular versions of these models when applied `out-of-the-box' for downstream tasks. We focus on generative language models as they are well-suited for extracting biases inherited from training data. Specifically, we conduct an in-depth analysis of GPT-2, which is the most downloaded text generation model on HuggingFace, with over half a million downloads per month. We assess biases related to occupational associations for different protected categories by intersecting gender with religion, sexuality, ethnicity, political affiliation, and continental name origin. Using a template-based data collection pipeline, we collect 396K sentence completions made by GPT-2 and find: (i) The machine-predicted jobs are less diverse and more stereotypical for women than for men, especially for intersections; (ii) Intersectional interactions are highly relevant for occupational associations, which we quantify by fitting 262 logistic models; (iii) For most occupations, GPT-2 reflects the skewed gender and ethnicity distribution found in US Labor Bureau data, and even pulls the societally-skewed distribution towards gender parity in cases where its predictions deviate from real labor market observations. This raises the normative question of what language models \textit{should} learn - whether they should reflect or correct for existing inequalities.
****
Weisfeiler and Lehman Go Cellular: CW Networks
Cristian Bodnar, Fabrizio Frasca, Nina Otter, Yuguang Wang, Pietro Liò, Guido F. Montufar, Michael Bronstein
Graph Neural Networks (GNNs) are limited in their expressive power, struggle with long-range interactions and lack a principled way to model higher-order structures. These problems can be attributed to the strong coupling between the computational graph and the input graph structure. The recently proposed Message Passing Simplicial Networks naturally decouple these elements by performing message passing on the clique complex of the graph. Nevertheless, these models can be severely constrained by the rigid combinatorial structure of Simplicial Complexes (SCs). In this work, we extend recent theoretical results on SCs to regular Cell Complexes, topological objects that flexibly subsume SCs and graphs. We show that this generalisation provides a powerful set of graph "lifting" transformations, each leading to a unique hierarchical message passing procedure. The resulting methods, which we collectively call CW Networks (CWNs), are strictly more powerful than the WL test and not less powerful than the 3-WL test. In particular, we demonstrate the effectiveness of one such scheme, based on rings, when applied to molecular graph problems. The proposed architecture benefits from provably larger expressivity than commonly used GNNs, principled modelling of higher-order signals and from compressing the distances between nodes. We demonstrate that our model achieves state-of-the-art results on a variety of molecular datasets.
****
Learning Conjoint Attentions for Graph Neural Nets
Tiantian He, Yew Soon Ong, L Bai
In this paper, we present Conjoint Attentions (CAs), a class of novel learning-to-attend strategies for graph neural networks (GNNs). Besides considering the layer-wise node features propagated within the GNN, CAs can additionally incorpora

te various structural interventions, such as node cluster embedding, and higher-order structural correlations that can be learned outside of GNN, when computing attention scores. The node features that are regarded as significant by the con joint criteria are therefore more likely to be propagated in the GNN. Given the novel Conjoint Attention strategies, we then propose Graph conjoint attention ne tworks (CATs) that can learn representations embedded with significant latent fe atures deemed by the Conjoint Attentions. Besides, we theoretically validate the discriminative capacity of CATs. CATs utilizing the proposed Conjoint Attentio n strategies have been extensively tested in well-established benchmarking datas ets and comprehensively compared with state-of-the-art baselines. The obtained n otable performance demonstrates the effectiveness of the proposed Conjoint Atten tions.
****
Hybrid Regret Bounds for Combinatorial Semi-Bandits and Adversarial Linear Bandi ts
Shinji Ito
This study aims to develop bandit algorithms that automatically exploit tendenci es of certain environments to improve performance, without any prior knowledge r egarding the environments. We first propose an algorithm for combinatorial semi-bandits with a hybrid regret bound that includes two main features: a best-of-th ree-worlds guarantee and multiple data-dependent regret bounds. The former means that the algorithm will work nearly optimally in all environments in an adversa rial setting, a stochastic setting, or a stochastic setting with adversarial cor ruptions. The latter implies that, even if the environment is far from exhibitin g stochastic behavior, the algorithm will perform better as long as the environm ent is "easy" in terms of certain metrics. The metrics w.r.t. the easiness refer red to in this paper include cumulative loss for optimal actions, total quadrati c variation of losses, and path-length of a loss sequence. We also show hybrid d ata-dependent regret bounds for adversarial linear bandits, which include a firs t path-length regret bound that is tight up to logarithmic factors.
****
Pay Better Attention to Attention: Head Selection in Multilingual and Multi-Doma in Sequence Modeling
Hongyu Gong, Yun Tang, Juan Pino, Xian 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.
****
Cardinality-Regularized Hawkes-Granger Model
Tsuyoshi Ide, Georgios Kollias, Dzung Phan, Naoki Abe
We propose a new sparse Granger-causal learning framework for temporal event dat a. We focus on a specific class of point processes called the Hawkes process. We begin by pointing out that most of the existing sparse causal learning algorith ms for the Hawkes process suffer from a singularity in maximum likelihood estima tion. As a result, their sparse solutions can appear only as numerical artifacts . In this paper, we propose a mathematically well-defined sparse causal learning framework based on a cardinality-regularized Hawkes process, which remedies the pathological issues of existing approaches. We leverage the proposed algorithm for the task of instance-wise causal event analysis, where sparsity plays a crit ical role. We validate the proposed framework with two real use-cases, one from the power grid and the other from the cloud data center management domain.
****
Aligned Structured Sparsity Learning for Efficient Image Super-Resolution
Yulun Zhang, Huan Wang, Can Qin, Yun Fu
Requests for name changes in the electronic proceedings will be accepted with 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.
****

Why Lottery Ticket Wins? A Theoretical Perspective of Sample Complexity on Sparse Neural Networks
Shuai Zhang, Meng Wang, Sijia Liu, Pin-Yu Chen, Jinjun Xiong
The lottery ticket hypothesis (LTH) states that learning on a properly pruned network (the winning ticket) has improved test accuracy over the original unpruned network. Although LTH has been justified empirically in a broad range of deep neural network (DNN) involved applications like computer vision and natural language processing, the theoretical validation of the improved generalization of a winning ticket remains elusive. To the best of our knowledge, our work, for the first time, characterizes the performance of training a pruned neural network by analyzing the geometric structure of the objective function and the sample complexity to achieve zero generalization error. We show that the convex region near a desirable model with guaranteed generalization enlarges as the neural network model is pruned, indicating the structural importance of a winning ticket. Moreover, as the algorithm for training a pruned neural network is specified as an (accelerated) stochastic gradient descent algorithm, we theoretically show that the number of samples required for achieving zero generalization error is proportional to the number of the non-pruned weights in the hidden layer. With a fixed number of samples, training a pruned neural network enjoys a faster convergence rate to the desired model than training the original unpruned one, providing a formal justification of the improved generalization of the winning ticket. Our theoretical results are acquired from learning a pruned neural network of one hidden layer, while experimental results are further provided to justify the implications in pruning multi-layer neural networks.
****
Constrained Robust Submodular Partitioning
Shengjie Wang, Tianyi Zhou, Chandrashekhar Lavania, Jeff A Bilmes
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Online Knapsack with Frequency Predictions
Sungjin Im, Ravi Kumar, Mahshid Montazer Qaem, Manish Purohit
There has been recent interest in using machine-learned predictions to improve the worst-case guarantees of online algorithms. In this paper we continue this line of work by studying the online knapsack problem, but with very weak predictions: in the form of knowing an upper and lower bound for the number of items of each value. We systematically derive online algorithms that attain the best possible competitive ratio for any fixed prediction; we also extend the results to more general settings such as generalized one-way trading and two-stage online knapsack. Our work shows that even seemingly weak predictions can be utilized effectively to provably improve the performance of online algorithms.
****
On Component Interactions in Two-Stage Recommender Systems
Jiri Hron, Karl Krauth, Michael Jordan, Niki Kilbertus
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Lip to Speech Synthesis with Visual Context Attentional GAN
Minsu Kim, Joanna Hong, Yong Man Ro
In this paper, we propose a novel lip-to-speech generative adversarial network, Visual Context Attentional GAN (VCA-GAN), which can jointly model local and global lip movements during speech synthesis. Specifically, the proposed VCA-GAN synthesizes the speech from local lip visual features by finding a mapping function of viseme-to-phoneme, while global visual context is embedded into the intermediate layers of the generator to clarify the ambiguity in the mapping induced by homophene. To achieve this, a visual context attention module is proposed where

it encodes global representations from the local visual features, and provides the desired global visual context corresponding to the given coarse speech representation to the generator through audio-visual attention. In addition to the explicit modelling of local and global visual representations, synchronization learning is introduced as a form of contrastive learning that guides the generator to synthesize a speech in sync with the given input lip movements. Extensive experiments demonstrate that the proposed VCA-GAN outperforms existing state-of-the-art and is able to effectively synthesize the speech from multi-speaker that has been barely handled in the previous works.
****

Non-convex Distributionally Robust Optimization: Non-asymptotic Analysis
Jikai Jin, Bohang Zhang, Haiyang Wang, 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.
****

Goal-Aware Cross-Entropy for Multi-Target Reinforcement Learning
Kibeom Kim, Min Whoo Lee, Yoonsung Kim, JeHwan Ryu, Minsu Lee, Byoung-Tak Zhang
Learning in a multi-target environment without prior knowledge about the targets requires a large amount of samples and makes generalization difficult. To solve this problem, it is important to be able to discriminate targets through semantic understanding. In this paper, we propose goal-aware cross-entropy (GACE) loss, that can be utilized in a self-supervised way using auto-labeled goal states alongside reinforcement learning. Based on the loss, we then devise goal-discriminative attention networks (GDAN) which utilize the goal-relevant information to focus on the given instruction. We evaluate the proposed methods on visual navigation and robot arm manipulation tasks with multi-target environments and show that GDAN outperforms the state-of-the-art methods in terms of task success ratio, sample efficiency, and generalization. Additionally, qualitative analyses demonstrate that our proposed method can help the agent become aware of and focus on the given instruction clearly, promoting goal-directed behavior.
****

Smooth Normalizing Flows
Jonas Köhler, Andreas Krämer, Frank Noe
Normalizing flows are a promising tool for modeling probability distributions in physical systems. While state-of-the-art flows accurately approximate distributions and energies, applications in physics additionally require smooth energies to compute forces and higher-order derivatives. Furthermore, such densities are often defined on non-trivial topologies. A recent example are Boltzmann Generators for generating 3D-structures of peptides and small proteins. These generative models leverage the space of internal coordinates (dihedrals, angles, and bonds), which is a product of hypertori and compact intervals. In this work, we introduce a class of smooth mixture transformations working on both compact intervals and hypertori.Mixture transformations employ root-finding methods to invert them in practice, which has so far prevented bi-directional flow training. To this end, we show that parameter gradients and forces of such inverses can be computed from forward evaluations via the inverse function theorem.We demonstrate two advantages of such smooth flows: they allow training by force matching to simulation data and can be used as potentials in molecular dynamics simulations.
****

MetaAvatar: Learning Animatable Clothed Human Models from Few Depth Images
Shaofei Wang, Marko Mihajlovic, Qianli Ma, Andreas Geiger, Siyu Tang
In this paper, we aim to create generalizable and controllable neural signed distance fields (SDFs) that represent clothed humans from monocular depth observations. Recent advances in deep learning, especially neural implicit representations, have enabled human shape reconstruction and controllable avatar generation from different sensor inputs. However, to generate realistic cloth deformations from novel input poses, watertight meshes or dense full-body scans are usually needed as inputs. Furthermore, due to the difficulty of effectively modeling pose-d

ependent cloth deformations for diverse body shapes and cloth types, existing approaches resort to per-subject/cloth-type optimization from scratch, which is computationally expensive. In contrast, we propose an approach that can quickly generate realistic clothed human avatars, represented as controllable neural SDFs, given only monocular depth images. We achieve this by using meta-learning to learn an initialization of a hypernetwork that predicts the parameters of neural SDFs. The hypernetwork is conditioned on human poses and represents a clothed neural avatar that deforms non-rigidly according to the input poses. Meanwhile, it is meta-learned to effectively incorporate priors of diverse body shapes and cloth types and thus can be much faster to fine-tune, compared to models trained from scratch. We qualitatively and quantitatively show that our approach outperforms state-of-the-art approaches that require complete meshes as inputs while our approach requires only depth frames as inputs and runs orders of magnitudes faster. Furthermore, we demonstrate that our meta-learned hypernetwork is very robust, being the first to generate avatars with realistic dynamic cloth deformations given as few as 8 monocular depth frames.
****

Distributed Principal Component Analysis with Limited Communication
Foivos Alimisis, Peter Davies, Bart Vandereycken, Dan Alistarh
We study efficient distributed algorithms for the fundamental problem of principal component analysis and leading eigenvector computation on the sphere, when the data are randomly distributed among a set of computational nodes. We propose a new quantized variant of Riemannian gradient descent to solve this problem, and prove that the algorithm converges with high probability under a set of necessary spherical-convexity properties. We give bounds on the number of bits transmitted by the algorithm under common initialization schemes, and investigate the dependency on the problem dimension in each case.
****

Newton-LESS: Sparsification without Trade-offs for the Sketched Newton Update
Michal Derezinski, Jonathan Lacotte, Mert Pilanci, Michael W. Mahoney
In second-order optimization, a potential bottleneck can be computing the Hessian matrix of the optimized function at every iteration. Randomized sketching has emerged as a powerful technique for constructing estimates of the Hessian which can be used to perform approximate Newton steps. This involves multiplication by a random sketching matrix, which introduces a trade-off between the computational cost of sketching and the convergence rate of the optimization. A theoretically desirable but practically much too expensive choice is to use a dense Gaussian sketching matrix, which produces unbiased estimates of the exact Newton step and offers strong problem-independent convergence guarantees. We show that the Gaussian matrix can be drastically sparsified, substantially reducing the computational cost, without affecting its convergence properties in any way. This approach, called Newton-LESS, is based on a recently introduced sketching technique: LEverage Score Sparsified (LESS) embeddings. We prove that Newton-LESS enjoys nearly the same problem-independent local convergence rate as Gaussian embeddings for a large class of functions. In particular, this leads to a new state-of-the-art convergence result for an iterative least squares solver. Finally, we substantially extend LESS embeddings to include uniformly sparsified random sign matrices which can be implemented efficiently and perform well in numerical experiments.
****

Confident Anchor-Induced Multi-Source Free Domain Adaptation
Jiahua Dong, Zhen Fang, Anjin Liu, Gan Sun, Tongliang Liu
Unsupervised domain adaptation has attracted appealing academic attentions by transferring knowledge from labeled source domain to unlabeled target domain. However, most existing methods assume the source data are drawn from a single domain, which cannot be successfully applied to explore complementarily transferable knowledge from multiple source domains with large distribution discrepancies. Moreover, they require access to source data during training, which are inefficient and unpractical due to privacy preservation and memory storage. To address these challenges, we develop a novel Confident-Anchor-induced multi-source-free Doma

in Adaptation (CAiDA) model, which is a pioneer exploration of knowledge adaptation from multiple source domains to the unlabeled target domain without any source data, but with only pre-trained source models. Specifically, a source-specific transferable perception module is proposed to automatically quantify the contributions of the complementary knowledge transferred from multi-source domains to the target domain. To generate pseudo labels for the target domain without access to the source data, we develop a confident-anchor-induced pseudo label generator by constructing a confident anchor group and assigning each unconfident target sample with a semantic-nearest confident anchor. Furthermore, a class-relationship-aware consistency loss is proposed to preserve consistent inter-class relationships by aligning soft confusion matrices across domains. Theoretical analysis answers why multi-source domains are better than a single source domain, and establishes a novel learning bound to show the effectiveness of exploiting multi-source domains. Experiments on several representative datasets illustrate the superiority of our proposed CAiDA model. The code is available at https://github.com/Learning-group123/CAiDA.
****

Word2Fun: Modelling Words as Functions for Diachronic Word Representation
Benyou Wang, Emanuele Di Buccio, Massimo Melucci
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

Iteratively Reweighted Least Squares for Basis Pursuit with Global Linear Convergence Rate
Christian Kümmerle, Claudio Mayrink Verdun, Dominik Stöger
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

Low-Rank Constraints for Fast Inference in Structured Models
Justin Chiu, Yuntian Deng, Alexander Rush
Structured distributions, i.e. distributions over combinatorial spaces, are commonly used to learn latent probabilistic representations from observed data. However, scaling these models is bottlenecked by the high computational and memory complexity with respect to the size of the latent representations. Common models such as Hidden Markov Models (HMMs) and Probabilistic Context-Free Grammars (PCFGs) require time and space quadratic and cubic in the number of hidden states respectively. This work demonstrates a simple approach to reduce the computational and memory complexity of a large class of structured models. We show that by viewing the central inference step as a matrix-vector product and using a low-rank constraint, we can trade off model expressivity and speed via the rank. Experiments with neural parameterized structured models for language modeling, polyphonic music modeling, unsupervised grammar induction, and video modeling show that our approach matches the accuracy of standard models at large state spaces while providing practical speedups.
****

Accumulative Poisoning Attacks on Real-time Data
Tianyu Pang, Xiao Yang, Yinpeng Dong, Hang Su, Jun Zhu
Collecting training data from untrusted sources exposes machine learning services to poisoning adversaries, who maliciously manipulate training data to degrade the model accuracy. When trained on offline datasets, poisoning adversaries have to inject the poisoned data in advance before training, and the order of feeding these poisoned batches into the model is stochastic. In contrast, practical systems are more usually trained/fine-tuned on sequentially captured real-time data, in which case poisoning adversaries could dynamically poison each data batch according to the current model state. In this paper, we focus on the real-time settings and propose a new attacking strategy, which affiliates an accumulative p

hase with poisoning attacks to secretly (i.e., without affecting accuracy) magni
fy the destructive effect of a (poisoned) trigger batch. By mimicking online lea
rning and federated learning on MNIST and CIFAR-10, we show that model accuracy
significantly drops by a single update step on the trigger batch after the accum
ulative phase. Our work validates that a well-designed but straightforward attac
king strategy can dramatically amplify the poisoning effects, with no need to ex
plore complex techniques.
****
UCB-based Algorithms for Multinomial Logistic Regression Bandits
Sanae Amani, Christos Thrampoulidis
Requests for name changes in the electronic proceedings will be accepted with 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 the Long-Term Effects of Novel Treatments
Keith Battocchi, Eleanor Dillon, Maggie Hei, Greg Lewis, Miruna Oprescu, Vasilis
 Syrgkanis
Policy makers often need to estimate the long-term effects of novel treatments,
while only having historical data of older treatment options. We propose a surro
gate-based approach using a long-term dataset where only past treatments were ad
ministered and a short-term dataset where novel treatments have been administere
d. Our approach generalizes previous surrogate-style methods, allowing for conti
nuous treatments and serially-correlated treatment policies while maintaining co
nsistency and root-n asymptotically normal estimates under a Markovian assumptio
n on the data and the observational policy. Using a semi-synthetic dataset on cu
stomer incentives from a major corporation, we evaluate the performance of our m
ethod and discuss solutions to practical challenges when deploying our methodolo
gy.
****
Dual Progressive Prototype Network for Generalized Zero-Shot Learning
Chaoqun Wang, Shaobo Min, Xuejin Chen, Xiaoyan Sun, Houqiang Li
Generalized Zero-Shot Learning (GZSL) aims to recognize new categories with auxi
liary semantic information, e.g., category attributes. In this paper, we handle
the critical issue of domain shift problem, i.e., confusion between seen and uns
een categories, by progressively improving cross-domain transferability and cate
gory discriminability of visual representations. Our approach, named Dual Progre
ssive Prototype Network (DPPN), constructs two types of prototypes that record p
rototypical visual patterns for attributes and categories, respectively. With at
tribute prototypes, DPPN alternately searches attribute-related local regions an
d updates corresponding attribute prototypes to progressively explore accurate a
ttribute-region correspondence. This enables DPPN to produce visual representati
ons with accurate attribute localization ability, which benefits the semantic-vi
sual alignment and representation transferability. Besides, along with progressi
ve attribute localization, DPPN further projects category prototypes into multip
le spaces to progressively repel visual representations from different categorie
s, which boosts category discriminability. Both attribute and category prototype
s are collaboratively learned in a unified framework, which makes visual represe
ntations of DPPN transferable and distinctive.Experiments on four benchmarks pro
ve that DPPN effectively alleviates the domain shift problem in GZSL.
****
Derivative-Free Policy Optimization for Linear Risk-Sensitive and Robust Control
 Design: Implicit Regularization and Sample Complexity
Kaiqing Zhang, Xiangyuan Zhang, Bin Hu, Tamer Basar
Direct policy search serves as one of the workhorses in modern reinforcement lea
rning (RL), and its applications in continuous control tasks have recently attra
cted increasing attention. In this work, we investigate the convergence theory o
f policy gradient (PG) methods for learning the linear risk-sensitive and robust
 controller. In particular, we develop PG methods that can be implemented in a d
erivative-free fashion by sampling system trajectories, and establish both globa

l convergence and sample complexity results in the solutions of two fundamental settings in risk-sensitive and robust control: the finite-horizon linear exponen tial quadratic Gaussian, and the finite-horizon linear-quadratic disturbance att enuation problems. As a by-product, our results also provide the first sample co mplexity for the global convergence of PG methods on solving zero-sum linear-qua dratic dynamic games, a nonconvex-nonconcave minimax optimization problem that s erves as a baseline setting in multi-agent reinforcement learning (MARL) with co ntinuous spaces. One feature of our algorithms is that during the learning phase , a certain level of robustness/risk-sensitivity of the controller is preserved, which we termed as the implicit regularization property, and is an essential re quirement in safety-critical control systems.
****
G-PATE: Scalable Differentially Private Data Generator via Private Aggregation o f Teacher Discriminators
Yunhui Long, Boxin Wang, Zhuolin Yang, Bhavya Kailkhura, Aston Zhang, Carl Gunte r, Bo 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.
****
On the Existence of The Adversarial Bayes Classifier
Pranjal Awasthi, Natalie Frank, Mehryar Mohri
Adversarial robustness is a critical property in a variety of modern machine lea rning applications. While it has been the subject of several recent theoretical studies, many important questions related to adversarial robustness are still op en. In this work, we study a fundamental question regarding Bayes optimality fo r adversarial robustness. We provide general sufficient conditions under which t he existence of a Bayes optimal classifier can be guaranteed for adversarial rob ustness. Our results can provide a useful tool for a subsequent study of surroga te losses in adversarial robustness and their consistency properties.
****
Convex-Concave Min-Max Stackelberg Games
Denizalp Goktas, Amy Greenwald
Min-max optimization problems (i.e., min-max games) have been attracting a great deal of attention because of their applicability to a wide range of machine lea rning problems. Although significant progress has been made recently, the litera ture to date has focused on games with independent strategy sets; little is know n about solving games with dependent strategy sets, which can be characterized a s min-max Stackelberg games. We introduce two first-order methods that solve a l arge class of convex-concave min-max Stackelberg games, and show that our method s converge in polynomial time. Min-max Stackelberg games were first studied by W ald, under the posthumous name of Wald’s maximin model, a variant of which is th e main paradigm used in robust optimization, which means that our methods can li kewise solve many convex robust optimization problems. We observe that the compu tation of competitive equilibria in Fisher markets also comprises a min-max Stac kelberg game. Further, we demonstrate the efficacy and efficiency of our algori thms in practice by computing competitive equilibria in Fisher markets with vary ing utility structures. Our experiments suggest potential ways to extend our the oretical results, by demonstrating how different smoothness properties can affec t the convergence rate of our algorithms.
****
Misspecified Gaussian Process Bandit Optimization
Ilija Bogunovic, 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-auth ors prior to requesting a name change in the electronic proceedings.
****
Visual Adversarial Imitation Learning using Variational Models

Rafael Rafailov, Tianhe Yu, Aravind Rajeswaran, Chelsea Finn
Reward function specification, which requires considerable human effort and iter
ation, remains a major impediment for learning behaviors through deep reinforcem
ent learning. In contrast, providing visual demonstrations of desired behaviors
presents an easier and more natural way to teach agents. We consider a setting w
here an agent is provided a fixed dataset of visual demonstrations illustrating
how to perform a task, and must learn to solve the task using the provided demon
strations and unsupervised environment interactions. This setting presents a num
ber of challenges including representation learning for visual observations, sam
ple complexity due to high dimensional spaces, and learning instability due to t
he lack of a fixed reward or learning signal. Towards addressing these challenge
s, we develop a variational model-based adversarial imitation learning (V-MAIL)
algorithm. The model-based approach provides a strong signal for representation
learning, enables sample efficiency, and improves the stability of adversarial t
raining by enabling on-policy learning. Through experiments involving several vi
sion-based locomotion and manipulation tasks, we find that V-MAIL learns success
ful visuomotor policies in a sample-efficient manner, has better stability compa
red to prior work, and also achieves higher asymptotic performance. We further f
ind that by transferring the learned models, V-MAIL can learn new tasks from vis
ual demonstrations without any additional environment interactions. All results
including videos can be found online at https://sites.google.com/view/variationa
l-mail
****
Object-Aware Regularization for Addressing Causal Confusion in Imitation Learnin
g
Jongjin Park, Younggyo Seo, Chang Liu, Li Zhao, Tao Qin, Jinwoo Shin, Tie-Yan Li
u
Behavioral cloning has proven to be effective for learning sequential decision-m
aking policies from expert demonstrations. However, behavioral cloning often suf
fers from the causal confusion problem where a policy relies on the noticeable e
ffect of expert actions due to the strong correlation but not the cause we desir
e. This paper presents Object-aware REgularizatiOn (OREO), a simple technique th
at regularizes an imitation policy in an object-aware manner. Our main idea is t
o encourage a policy to uniformly attend to all semantic objects, in order to pr
event the policy from exploiting nuisance variables strongly correlated with exp
ert actions. To this end, we introduce a two-stage approach: (a) we extract sema
ntic objects from images by utilizing discrete codes from a vector-quantized var
iational autoencoder, and (b) we randomly drop the units that share the same dis
crete code together, i.e., masking out semantic objects. Our experiments demonst
rate that OREO significantly improves the performance of behavioral cloning, out
performing various other regularization and causality-based methods on a variety
 of Atari environments and a self-driving CARLA environment. We also show that o
ur method even outperforms inverse reinforcement learning methods trained with a
 considerable amount of environment interaction.
****
Reliable and Trustworthy Machine Learning for Health Using Dataset Shift Detecti
on
Chunjong Park, Anas Awadalla, Tadayoshi Kohno, Shwetak Patel
Unpredictable ML model behavior on unseen data, especially in the health domain,
 raises serious concerns about its safety as repercussions for mistakes can be f
atal. In this paper, we explore the feasibility of using state-of-the-art out-of
-distribution detectors for reliable and trustworthy diagnostic predictions. We
select publicly available deep learning models relating to various health condit
ions (e.g., skin cancer, lung sound, and Parkinson's disease) using various inpu
t data types (e.g., image, audio, and motion data). We demonstrate that these mo
dels show unreasonable predictions on out-of-distribution datasets. We show that
 Mahalanobis distance- and Gram matrices-based out-of-distribution detection met
hods are able to detect out-of-distribution data with high accuracy for the heal
th models that operate on different modalities. We then translate the out-of-dis
tribution score into a human interpretable \textsc{confidence score} to investig

ate its effect on the users' interaction with health ML applications. Our user study shows that the \textsc{confidence score} helped the participants only trust the results with a high score to make a medical decision and disregard results with a low score. Through this work, we demonstrate that dataset shift is a critical piece of information for high-stake ML applications, such as medical diagnosis and healthcare, to provide reliable and trustworthy predictions to the users.
****
Multiclass Boosting and the Cost of Weak Learning
Nataly Brukhim, Elad Hazan, Shay Moran, Indraneel Mukherjee, Robert E. Schapire
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Partition-Based Formulations for Mixed-Integer Optimization of Trained ReLU Neural Networks
Calvin Tsay, Jan Kronqvist, Alexander Thebelt, Ruth Misener
This paper introduces a class of mixed-integer formulations for trained ReLU neural networks. The approach balances model size and tightness by partitioning node inputs into a number of groups and forming the convex hull over the partitions via disjunctive programming. At one extreme, one partition per input recovers the convex hull of a node, i.e., the tightest possible formulation for each node. For fewer partitions, we develop smaller relaxations that approximate the convex hull, and show that they outperform existing formulations. Specifically, we propose strategies for partitioning variables based on theoretical motivations and validate these strategies using extensive computational experiments. Furthermore, the proposed scheme complements known algorithmic approaches, e.g., optimization-based bound tightening captures dependencies within a partition.
****
Hyperparameter Optimization Is Deceiving Us, and How to Stop It
A. Feder Cooper, Yucheng Lu, Jessica Forde, Christopher M. De Sa
Recent empirical work shows that inconsistent results based on choice of hyperparameter optimization (HPO) configuration are a widespread problem in ML research. When comparing two algorithms J and K searching one subspace can yield the conclusion that J outperforms K, whereas searching another can entail the opposite. In short, the way we choose hyperparameters can deceive us. We provide a theoretical complement to this prior work, arguing that, to avoid such deception, the process of drawing conclusions from HPO should be made more rigorous. We call this process epistemic hyperparameter optimization (EHPO), and put forth a logical framework to capture its semantics and how it can lead to inconsistent conclusions about performance. Our framework enables us to prove EHPO methods that are guaranteed to be defended against deception, given bounded compute time budget t. We demonstrate our framework's utility by proving and empirically validating a defended variant of random search.
****
On the Convergence Theory of Debiased Model-Agnostic Meta-Reinforcement Learning
Alireza Fallah, Kristian Georgiev, Aryan Mokhtari, Asuman Ozdaglar
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
3D Pose Transfer with Correspondence Learning and Mesh Refinement
Chaoyue Song, Jiacheng Wei, Ruibo Li, Fayao Liu, Guosheng Lin
3D pose transfer is one of the most challenging 3D generation tasks. It aims to transfer the pose of a source mesh to a target mesh and keep the identity (e.g., body shape) of the target mesh. Some previous works require key point annotations to build reliable correspondence between the source and target meshes, while other methods do not consider any shape correspondence between sources and targe

ts, which leads to limited generation quality. In this work, we propose a correspondence-refinement network to achieve the 3D pose transfer for both human and animal meshes. The correspondence between source and target meshes is first established by solving an optimal transport problem. Then, we warp the source mesh according to the dense correspondence and obtain a coarse warped mesh. The warped mesh will be better refined with our proposed Elastic Instance Normalization, which is a conditional normalization layer and can help to generate high-quality meshes. Extensive experimental results show that the proposed architecture can effectively transfer the poses from source to target meshes and produce better results with satisfied visual performance than state-of-the-art methods.
****

Framing RNN as a kernel method: A neural ODE approach
Adeline Fermanian, Pierre Marion, Jean-Philippe Vert, Gérard Biau
Building on the interpretation of a recurrent neural network (RNN) as a continuous-time neural differential equation, we show, under appropriate conditions, that the solution of a RNN can be viewed as a linear function of a specific feature set of the input sequence, known as the signature. This connection allows us to frame a RNN as a kernel method in a suitable reproducing kernel Hilbert space. As a consequence, we obtain theoretical guarantees on generalization and stability for a large class of recurrent networks. Our results are illustrated on simulated datasets.
****

Contextual Similarity Aggregation with Self-attention for Visual Re-ranking
Jianbo Ouyang, Hui Wu, Min Wang, Wengang Zhou, Houqiang Li
In content-based image retrieval, the first-round retrieval result by simple visual feature comparison may be unsatisfactory, which can be refined by visual re-ranking techniques. In image retrieval, it is observed that the contextual similarity among the top-ranked images is an important clue to distinguish the semantic relevance. Inspired by this observation, in this paper, we propose a visual re-ranking method by contextual similarity aggregation with self-attention. In our approach, for each image in the top-K ranking list, we represent it into an affinity feature vector by comparing it with a set of anchor images. Then, the affinity features of the top-K images are refined by aggregating the contextual information with a transformer encoder. Finally, the affinity features are used to recalculate the similarity scores between the query and the top-K images for re-ranking of the latter. To further improve the robustness of our re-ranking model and enhance the performance of our method, a new data augmentation scheme is designed. Since our re-ranking model is not directly involved with the visual feature used in the initial retrieval, it is ready to be applied to retrieval result lists obtained from various retrieval algorithms. We conduct comprehensive experiments on four benchmark datasets to demonstrate the generality and effectiveness of our proposed visual re-ranking method.
****

Can Information Flows Suggest Targets for Interventions in Neural Circuits?
Praveen Venkatesh, Sanghamitra Dutta, Neil Mehta, Pulkit Grover
Motivated by neuroscientific and clinical applications, we empirically examine whether observational measures of information flow can suggest interventions. We do so by performing experiments on artificial neural networks in the context of fairness in machine learning, where the goal is to induce fairness in the system through interventions. Using our recently developed M-information flow framework, we measure the flow of information about the true label (responsible for accuracy, and hence desirable), and separately, the flow of information about a protected attribute (responsible for bias, and hence undesirable) on the edges of a trained neural network. We then compare the flow magnitudes against the effect of intervening on those edges by pruning. We show that pruning edges that carry larger information flows about the protected attribute reduces bias at the output to a greater extent. This demonstrates that M-information flow can meaningfully suggest targets for interventions, answering the title's question in the affirmative. We also evaluate bias-accuracy tradeoffs for different intervention strategies, to analyze how one might use estimates of desirable and undesirable infor

mation flows (here, accuracy and bias flows) to inform interventions that preser
ve the former while reducing the latter.
****
AutoBalance: Optimized Loss Functions for Imbalanced Data
Mingchen Li, Xuechen Zhang, Christos Thrampoulidis, Jiasi Chen, Samet Oymak
Imbalanced datasets are commonplace in modern machine learning problems. The pre
sence of under-represented classes or groups with sensitive attributes results i
n concerns about generalization and fairness. Such concerns are further exacerba
ted by the fact that large capacity deep nets can perfectly fit the training dat
a and appear to achieve perfect accuracy and fairness during training, but perfo
rm poorly during test. To address these challenges, we propose AutoBalance, a bi
-level optimization framework that automatically designs a training loss functio
n to optimize a blend of accuracy and fairness-seeking objectives. Specifically,
 a lower-level problem trains the model weights, and an upper-level problem tune
s the loss function by monitoring and optimizing the desired objective over the
validation data. Our loss design enables personalized treatment for classes/grou
ps by employing a parametric cross-entropy loss and individualized data augmenta
tion schemes. We evaluate the benefits and performance of our approach for the a
pplication scenarios of imbalanced and group-sensitive classification. Extensive
 empirical evaluations demonstrate the benefits of AutoBalance over state-of-the
-art approaches. Our experimental findings are complemented with theoretical ins
ights on loss function design and the benefits of the train-validation split. Al
l code is available open-source.
****
SyncTwin: Treatment Effect Estimation with Longitudinal Outcomes
Zhaozhi Qian, Yao Zhang, Ioana Bica, Angela Wood, Mihaela van der Schaar
Most of the medical observational studies estimate the causal treatment effects
using electronic health records (EHR), where a patient's covariates and outcomes
 are both observed longitudinally. However, previous methods focus only on adjus
ting for the covariates while neglecting the temporal structure in the outcomes.
 To bridge the gap, this paper develops a new method, SyncTwin, that learns a pa
tient-specific time-constant representation from the pre-treatment observations.
 SyncTwin issues counterfactual prediction of a target patient by constructing a
 synthetic twin that closely matches the target in representation. The reliabili
ty of the estimated treatment effect can be assessed by comparing the observed a
nd synthetic pre-treatment outcomes. The medical experts can interpret the estim
ate by examining the most important contributing individuals to the synthetic tw
in. In the real-data experiment, SyncTwin successfully reproduced the findings o
f a randomized controlled clinical trial using observational data, which demonst
rates its usability in the complex real-world EHR.
****
Statistical Query Lower Bounds for List-Decodable Linear Regression
Ilias Diakonikolas, Daniel Kane, Ankit Pensia, Thanasis Pittas, Alistair Stewart
Requests for name changes in the electronic proceedings will be accepted with no
 questions asked. However name changes may cause bibliographic tracking issues.
  Authors are asked to consider this carefully and discuss it with their co-auth
ors prior to requesting a name change in the electronic proceedings.
****
Unsupervised Motion Representation Learning with Capsule Autoencoders
Ziwei Xu, Xudong Shen, Yongkang Wong, Mohan S. Kankanhalli
We propose the Motion Capsule Autoencoder (MCAE), which addresses a key challeng
e in the unsupervised learning of motion representations: transformation invaria
nce. MCAE models motion in a two-level hierarchy. In the lower level, a spatio-t
emporal motion signal is divided into short, local, and semantic-agnostic snippe
ts. In the higher level, the snippets are aggregated to form full-length semanti
c-aware segments. For both levels, we represent motion with a set of learned tra
nsformation invariant templates and the corresponding geometric transformations
by using capsule autoencoders of a novel design. This leads to a robust and effi
cient encoding of viewpoint changes. MCAE is evaluated on a novel Trajectory20 m
otion dataset and various real-world skeleton-based human action datasets. Notab

ly, it achieves better results than baselines on Trajectory20 with considerably fewer parameters and state-of-the-art performance on the unsupervised skeleton-based action recognition task.
****
VigDet: Knowledge Informed Neural Temporal Point Process for Coordination Detection on Social Media
Yizhou Zhang, Karishma Sharma, Yan Liu
Recent years have witnessed an increasing use of coordinated accounts on social media, operated by misinformation campaigns to influence public opinion and manipulate social outcomes. Consequently, there is an urgent need to develop an effective methodology for coordinated group detection to combat the misinformation on social media. However, existing works suffer from various drawbacks, such as, either limited performance due to extreme reliance on predefined signatures of coordination, or instead an inability to address the natural sparsity of account activities on social media with useful prior domain knowledge. Therefore, in this paper, we propose a coordination detection framework incorporating neural temporal point process with prior knowledge such as temporal logic or pre-defined filtering functions. Specifically, when modeling the observed data from social media with neural temporal point process, we jointly learn a Gibbs-like distribution of group assignment based on how consistent an assignment is to (1) the account embedding space and (2) the prior knowledge. To address the challenge that the distribution is hard to be efficiently computed and sampled from, we design a theoretically guaranteed variational inference approach to learn a mean-field approximation for it. Experimental results on a real-world dataset show the effectiveness of our proposed method compared to the SOTA model in both unsupervised and semi-supervised settings. We further apply our model on a COVID-19 Vaccine Tweets dataset. The detection result suggests the presence of suspicious coordinated efforts on spreading misinformation about COVID-19 vaccines.
****
An Improved Analysis and Rates for Variance Reduction under Without-replacement Sampling Orders
Xinmeng Huang, Kun Yuan, Xianghui Mao, Wotao Yin
When applying a stochastic algorithm, one must choose an order to draw samples. The practical choices are without-replacement sampling orders, which are empirically faster and more cache-friendly than uniform-iid-sampling but often have inferior theoretical guarantees. Without-replacement sampling is well understood only for SGD without variance reduction. In this paper, we will improve the convergence analysis and rates of variance reduction under without-replacement sampling orders for composite finite-sum minimization.Our results are in two-folds. First, we develop a damped variant of Finito called Prox-DFinito and establish its convergence rates with random reshuffling, cyclic sampling, and shuffling-once, under both generally and strongly convex scenarios. These rates match full-batch gradient descent and are state-of-the-art compared to the existing results for without-replacement sampling with variance-reduction. Second, our analysis can gauge how the cyclic order will influence the rate of cyclic sampling and, thus, allows us to derive the optimal fixed ordering. In the highly data-heterogeneous scenario, Prox-DFinito with optimal cyclic sampling can attain a sample-size-independent convergence rate, which, to our knowledge, is the first result that can match with uniform-iid-sampling with variance reduction. We also propose a practical method to discover the optimal cyclic ordering numerically.
****
Exploring Forensic Dental Identification with Deep Learning
Yuan Liang, Weikun Han, Liang Qiu, Chen Wu, Yiting Shao, Kun Wang, Lei He
Dental forensic identification targets to identify persons with dental traces.The task is vital for the investigation of criminal scenes and mass disasters because of the resistance of dental structures and the wide-existence of dental imaging. However, no widely accepted automated solution is available for this labour-costly task. In this work, we pioneer to study deep learning for dental forensic identification based on panoramic radiographs. We construct a comprehensive benchmark with various dental variations that can adequately reflect the difficult

ies of the task. By considering the task's unique challenges, we propose FoID, a deep learning method featured by: (\textit{i}) clinical-inspired attention localization, (\textit{ii}) domain-specific augmentations that enable instance discriminative learning, and (\textit{iii}) transformer-based self-attention mechanism that dynamically reasons the relative importance of attentions. We show that FoID can outperform traditional approaches by at least \textbf{22.98\%} in terms of Rank-1 accuracy, and outperform strong CNN baselines by at least \textbf{10.50\%} in terms of mean Average Precision (mAP). Moreover, extensive ablation studies verify the effectiveness of each building blocks of FoID. Our work can be a first step towards the automated system for forensic identification among large-scale multi-site databases. Also, the proposed techniques, \textit{e.g.}, self-attention mechanism, can also be meaningful for other identification tasks, \textit{e.g.}, pedestrian re-identification.Related data and codes can be found at \href{https://github.com/liangyuandg/FoID}{https://github.com/liangyuandg/FoID}.
****

Learning to Generate Realistic Noisy Images via Pixel-level Noise-aware Adversarial Training

Yuanhao Cai, Xiaowan Hu, Haoqian Wang, Yulun Zhang, Hanspeter Pfister, Donglai Wei

Existing deep learning real denoising methods require a large amount of noisy-clean image pairs for supervision. Nonetheless, capturing a real noisy-clean dataset is an unacceptable expensive and cumbersome procedure. To alleviate this problem, this work investigates how to generate realistic noisy images. Firstly, we formulate a simple yet reasonable noise model that treats each real noisy pixel as a random variable. This model splits the noisy image generation problem into two sub-problems: image domain alignment and noise domain alignment. Subsequently, we propose a novel framework, namely Pixel-level Noise-aware Generative Adversarial Network (PNGAN). PNGAN employs a pre-trained real denoiser to map the fake and real noisy images into a nearly noise-free solution space to perform image domain alignment. Simultaneously, PNGAN establishes a pixel-level adversarial training to conduct noise domain alignment. Additionally, for better noise fitting, we present an efficient architecture Simple Multi-scale Network (SMNet) as the generator. Qualitative validation shows that noise generated by PNGAN is highly similar to real noise in terms of intensity and distribution. Quantitative experiments demonstrate that a series of denoisers trained with the generated noisy images achieve state-of-the-art (SOTA) results on four real denoising benchmarks.
****

Multi-Agent Reinforcement Learning for Active Voltage Control on Power Distribution Networks

Jianhong Wang, Wangkun Xu, Yunjie Gu, Wenbin Song, Tim C Green

This paper presents a problem in power networks that creates an exciting and yet challenging real-world scenario for application of multi-agent reinforcement learning (MARL). The emerging trend of decarbonisation is placing excessive stress on power distribution networks. Active voltage control is seen as a promising solution to relieve power congestion and improve voltage quality without extra hardware investment, taking advantage of the controllable apparatuses in the network, such as roof-top photovoltaics (PVs) and static var compensators (SVCs). These controllable apparatuses appear in a vast number and are distributed in a wide geographic area, making MARL a natural candidate. This paper formulates the active voltage control problem in the framework of Dec-POMDP and establishes an open-source environment. It aims to bridge the gap between the power community and the MARL community and be a drive force towards real-world applications of MARL algorithms. Finally, we analyse the special characteristics of the active voltage control problems that cause challenges (e.g. interpretability) for state-of-the-art MARL approaches, and summarise the potential directions.
****

Looking Beyond Single Images for Contrastive Semantic Segmentation Learning

FEIHU ZHANG, Philip Torr, Rene Ranftl, Stephan Richter

We present an approach to contrastive representation learning for semantic segme

ntation. Our approach leverages the representational power of existing feature e xtractors to find corresponding regions across images. These cross-image corresp ondences are used as auxiliary labels to guide the pixel-level selection of pos itive and negative samples for more effective contrastive learning in semantic s egmentation. We show that auxiliary labels can be generated from a variety of fe ature extractors, ranging from image classification networks that have been trai ned using unsupervised contrastive learning to segmentation models that have bee n trained on a small amount of labeled data. We additionally introduce a novel m etric for rapidly judging the quality of a given auxiliary-labeling strategy, an d empirically analyze various factors that influence the performance of contrast ive learning for semantic segmentation. We demonstrate the effectiveness of our method both in the low-data as well as the high-data regime on various datasets. Our experiments show that contrastive learning with our auxiliary-labeling appr oach consistently boosts semantic segmentation accuracy when compared to standar d ImageNet pretraining and outperforms existing approaches of contrastive and se mi-supervised semantic segmentation.
****

A Constant Approximation Algorithm for Sequential Random-Order No-Substitution k -Median Clustering

Tom Hess, Michal Moshkovitz, Sivan Sabato

We study k-median clustering under the sequential no-substitution setting. In th is setting, a data stream is sequentially observed, and some of the points are s elected by the algorithm as cluster centers. However, a point can be selected as a center only immediately after it is observed, before observing the next point . In addition, a selected center cannot be substituted later. We give the first algorithm for this setting that obtains a constant approximation factor on the o ptimal cost under a random arrival order, an exponential improvement over previo us work. This is also the first constant approximation guarantee that holds with out any structural assumptions on the input data. Moreover, the number of select ed centers is only quasi-linear in k. Our algorithm and analysis are based on a careful cost estimation that avoids outliers, a new concept of a linear bin div ision, and a multi-scale approach to center selection.
****

Dangers of Bayesian Model Averaging under Covariate Shift

Pavel Izmailov, Patrick Nicholson, Sanae Lotfi, Andrew G. Wilson

Approximate Bayesian inference for neural networks is considered a robust altern ative to standard training, often providing good performance on out-of-distribut ion data. However, Bayesian neural networks (BNNs) with high-fidelity approximat e inference via full-batch Hamiltonian Monte Carlo achieve poor generalization u nder covariate shift, even underperforming classical estimation. We explain this surprising result, showing how a Bayesian model average can in fact be problema tic under covariate shift, particularly in cases where linear dependencies in th e input features cause a lack of posterior contraction. We additionally show why the same issue does not affect many approximate inference procedures, or classi cal maximum a-posteriori (MAP) training. Finally, we propose novel priors that i mprove the robustness of BNNs to many sources of covariate shift.
****

Learning Equilibria in Matching Markets from Bandit Feedback

Meena Jagadeesan, Alexander Wei, Yixin Wang, Michael Jordan, Jacob Steinhardt

Large-scale, two-sided matching platforms must find market outcomes that align w ith user preferences while simultaneously learning these preferences from data. But since preferences are inherently uncertain during learning, the classical no tion of stability (Gale and Shapley, 1962; Shapley and Shubik, 1971) is unattain able in these settings. To bridge this gap, we develop a framework and algorithm s for learning stable market outcomes under uncertainty. Our primary setting is matching with transferable utilities, where the platform both matches agents and sets monetary transfers between them. We design an incentive-aware learning obj ective that captures the distance of a market outcome from equilibrium. Using th is objective, we analyze the complexity of learning as a function of preference structure, casting learning as a stochastic multi-armed bandit problem. Algorith

mically, we show that "optimism in the face of uncertainty," the principle under lying many bandit algorithms, applies to a primal-dual formulation of matching w ith transfers and leads to near-optimal regret bounds. Our work takes a first st ep toward elucidating when and how stable matchings arise in large, data-driven marketplaces.
****
Towards Lower Bounds on the Depth of ReLU Neural Networks
Christoph Hertrich, Amitabh Basu, Marco Di Summa, Martin Skutella
We contribute to a better understanding of the class of functions that is repres ented by a neural network with ReLU activations and a given architecture. Using techniques from mixed-integer optimization, polyhedral theory, and tropical geom etry, we provide a mathematical counterbalance to the universal approximation th eorems which suggest that a single hidden layer is sufficient for learning tasks . In particular, we investigate whether the class of exactly representable funct ions strictly increases by adding more layers (with no restrictions on size). Th is problem has potential impact on algorithmic and statistical aspects because o f the insight it provides into the class of functions represented by neural hypo thesis classes. However, to the best of our knowledge, this question has not bee n investigated in the neural network literature. We also present upper bounds on the sizes of neural networks required to represent functions in these neural hy pothesis classes.
****
The Limitations of Large Width in Neural Networks: A Deep Gaussian Process Persp ective
Geoff Pleiss, John P. Cunningham
Large width limits have been a recent focus of deep learning research: modulo co mputational practicalities, do wider networks outperform narrower ones? Answerin g this question has been challenging, as conventional networks gain representati onal power with width, potentially masking any negative effects. Our analysis in this paper decouples capacity and width via the generalization of neural networ ks to Deep Gaussian Processes (Deep GP), a class of nonparametric hierarchical m odels that subsume neural nets. In doing so, we aim to understand how width affe cts (standard) neural networks once they have sufficient capacity for a given mo deling task. Our theoretical and empirical results on Deep GP suggest that large width can be detrimental to hierarchical models. Surprisingly, we prove that ev en nonparametric Deep GP converge to Gaussian processes, effectively becoming sh allower without any increase in representational power. The posterior, which cor responds to a mixture of data-adaptable basis functions, becomes less data-depen dent with width. Our tail analysis demonstrates that width and depth have opposi te effects: depth accentuates a model’s non-Gaussianity, while width makes model s increasingly Gaussian. We find there is a “sweet spot” that maximizes test per formance before the limiting GP behavior prevents adaptability, occurring at wid th = 1 or width = 2 for nonparametric Deep GP. These results make strong predict ions about the same phenomenon in conventional neural networks trained with L2 r egularization (analogous to a Gaussian prior on parameters): we show that such n eural networks may need up to 500 − 1000 hidden units for sufficient capacity - depending on the dataset - but further width degrades performance.
****
Exact marginal prior distributions of finite Bayesian neural networks
Jacob Zavatone-Veth, Cengiz Pehlevan
Requests for name changes in the electronic proceedings will be accepted with 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.
****
Spatiotemporal Joint Filter Decomposition in 3D Convolutional Neural Networks
Zichen Miao, Ze Wang, Xiuyuan Cheng, Qiang Qiu
In this paper, we introduce spatiotemporal joint filter decomposition to decoupl e spatial and temporal learning, while preserving spatiotemporal dependency in a video. A 3D convolutional filter is now jointly decomposed over a set of spatia

l and temporal filter atoms respectively. In this way, a 3D convolutional layer becomes three: a temporal atom layer, a spatial atom layer, and a joint coefficient layer, all three remaining convolutional. One obvious arithmetic manipulation allowed in our joint decomposition is to swap spatial or temporal atoms with a set of atoms that have the same number but different sizes, while keeping the remaining unchanged. For example, as shown later, we can now achieve tempo-invariance by simply dilating temporal atoms only. To illustrate this useful atom-swapping property, we further demonstrate how such a decomposition permits the direct learning of 3D CNNs with full-size videos through iterations of two consecutive sub-stages of learning: In the temporal stage, full-temporal downsampled-spatial data are used to learn temporal atoms and joint coefficients while fixing spatial atoms. In the spatial stage, full-spatial downsampled-temporal data are used for spatial atoms and joint coefficients while fixing temporal atoms. We show empirically on multiple action recognition datasets that, the decoupled spatiotemporal learning significantly reduces the model memory footprints, and allows deep 3D CNNs to model high-spatial long-temporal dependency with limited computational resources while delivering comparable performance.
****
Pooling by Sliced-Wasserstein Embedding
Navid Naderializadeh, Joseph F Comer, Reed Andrews, Heiko Hoffmann, Soheil Kolouri
Learning representations from sets has become increasingly important with many applications in point cloud processing, graph learning, image/video recognition, and object detection. We introduce a geometrically-interpretable and generic pooling mechanism for aggregating a set of features into a fixed-dimensional representation. In particular, we treat elements of a set as samples from a probability distribution and propose an end-to-end trainable Euclidean embedding for sliced-Wasserstein distance to learn from set-structured data effectively. We evaluate our proposed pooling method on a wide variety of set-structured data, including point-cloud, graph, and image classification tasks, and demonstrate that our proposed method provides superior performance over existing set representation learning approaches. Our code is available at https://github.com/navid-naderi/PSWE.
****
On the Theory of Reinforcement Learning with Once-per-Episode Feedback
Niladri Chatterji, Aldo Pacchiano, Peter Bartlett, Michael Jordan
We study a theory of reinforcement learning (RL) in which the learner receives binary feedback only once at the end of an episode. While this is an extreme test case for theory, it is also arguably more representative of real-world applications than the traditional requirement in RL practice that the learner receive feedback at every time step. Indeed, in many real-world applications of reinforcement learning, such as self-driving cars and robotics, it is easier to evaluate whether a learner's complete trajectory was either good'' orbad,'' but harder to provide a reward signal at each step. To show that learning is possible in this more challenging setting, we study the case where trajectory labels are generated by an unknown parametric model, and provide a statistically and computationally efficient algorithm that achieves sublinear regret.
****
ResNEsts and DenseNEsts: Block-based DNN Models with Improved Representation Guarantees
Kuan-Lin Chen, Ching-Hua Lee, Harinath Garudadri, Bhaskar D Rao
Models recently used in the literature proving residual networks (ResNets) are better than linear predictors are actually different from standard ResNets that have been widely used in computer vision. In addition to the assumptions such as scalar-valued output or single residual block, the models fundamentally considered in the literature have no nonlinearities at the final residual representation that feeds into the final affine layer. To codify such a difference in nonlinearities and reveal a linear estimation property, we define ResNEsts, i.e., Residual Nonlinear Estimators, by simply dropping nonlinearities at the last residual representation from standard ResNets. We show that wide ResNEsts with bottleneck

blocks can always guarantee a very desirable training property that standard ResNets aim to achieve, i.e., adding more blocks does not decrease performance given the same set of basis elements. To prove that, we first recognize ResNEsts are basis function models that are limited by a coupling problem in basis learning and linear prediction. Then, to decouple prediction weights from basis learning, we construct a special architecture termed augmented ResNEst (A-ResNEst) that always guarantees no worse performance with the addition of a block. As a result, such an A-ResNEst establishes empirical risk lower bounds for a ResNEst using corresponding bases. Our results demonstrate ResNEsts indeed have a problem of diminishing feature reuse; however, it can be avoided by sufficiently expanding or widening the input space, leading to the above-mentioned desirable property. Inspired by the densely connected networks (DenseNets) that have been shown to outperform ResNets, we also propose a corresponding new model called Densely connected Nonlinear Estimator (DenseNEst). We show that any DenseNEst can be represented as a wide ResNEst with bottleneck blocks. Unlike ResNEsts, DenseNEsts exhibit the desirable property without any special architectural re-design.
****
Locally private online change point detection
Tom Berrett, Yi 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.
****
Invariance Principle Meets Information Bottleneck for Out-of-Distribution Generalization
Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Jean-Christophe Gagnon-Audet, Yoshua Bengio, Ioannis Mitliagkas, Irina Rish
The invariance principle from causality is at the heart of notable approaches such as invariant risk minimization (IRM) that seek to address out-of-distribution (OOD) generalization failures. Despite the promising theory, invariance principle-based approaches fail in common classification tasks, where invariant (causal) features capture all the information about the label. Are these failures due to the methods failing to capture the invariance? Or is the invariance principle itself insufficient? To answer these questions, we revisit the fundamental assumptions in linear regression tasks, where invariance-based approaches were shown to provably generalize OOD. In contrast to the linear regression tasks, we show that for linear classification tasks we need much stronger restrictions on the distribution shifts, or otherwise OOD generalization is impossible. Furthermore, even with appropriate restrictions on distribution shifts in place, we show that the invariance principle alone is insufficient. We prove that a form of the information bottleneck constraint along with invariance helps address the key failures when invariant features capture all the information about the label and also retains the existing success when they do not. We propose an approach that incorporates both of these principles and demonstrate its effectiveness in several experiments.
****
Repulsive Deep Ensembles are Bayesian
Francesco D'Angelo, Vincent Fortuin
Deep ensembles have recently gained popularity in the deep learning community for their conceptual simplicity and efficiency. However, maintaining functional diversity between ensemble members that are independently trained with gradient descent is challenging. This can lead to pathologies when adding more ensemble members, such as a saturation of the ensemble performance, which converges to the performance of a single model. Moreover, this does not only affect the quality of its predictions, but even more so the uncertainty estimates of the ensemble, and thus its performance on out-of-distribution data. We hypothesize that this limitation can be overcome by discouraging different ensemble members from collapsing to the same function. To this end, we introduce a kernelized repulsive term in the update rule of the deep ensembles. We show that this simple modification n

ot only enforces and maintains diversity among the members but, even more import
antly, transforms the maximum a posteriori inference into proper Bayesian infere
nce. Namely, we show that the training dynamics of our proposed repulsive ensemb
les follow a Wasserstein gradient flow of the KL divergence with the true poster
ior. We study repulsive terms in weight and function space and empirically compa
re their performance to standard ensembles and Bayesian baselines on synthetic a
nd real-world prediction tasks.
****
BayesIMP: Uncertainty Quantification for Causal Data Fusion
Siu Lun Chau, Jean-Francois Ton, Javier González, Yee Teh, Dino Sejdinovic
While causal models are becoming one of the mainstays of machine learning, the p
roblem of uncertainty quantification in causal inference remains challenging. In
 this paper, we study the causal data fusion problem, where data arising from mu
ltiple causal graphs are combined to estimate the average treatment effect of a
target variable. As data arises from multiple sources and can vary in quality an
d sample size, principled uncertainty quantification becomes essential. To that
end, we introduce \emph{Bayesian Causal Mean Processes}, the framework which com
bines ideas from probabilistic integration and kernel mean embeddings to represe
nt interventional distributions in the reproducing kernel Hilbert space, while t
aking into account the uncertainty within each causal graph. To demonstrate the
informativeness of our uncertainty estimation, we apply our method to the Causal
 Bayesian Optimisation task and show improvements over state-of-the-art methods.
****
RMM: Reinforced Memory Management for Class-Incremental Learning
Yaoyao Liu, Bernt Schiele, Qianru Sun
Class-Incremental Learning (CIL) [38] trains classifiers under a strict memory b
udget: in each incremental phase, learning is done for new data, most of which i
s abandoned to free space for the next phase. The preserved data are exemplars u
sed for replaying. However, existing methods use a static and ad hoc strategy fo
r memory allocation, which is often sub-optimal. In this work, we propose a dyna
mic memory management strategy that is optimized for the incremental phases and
different object classes. We call our method reinforced memory management (RMM),
 leveraging reinforcement learning. RMM training is not naturally compatible wit
h CIL as the past, and future data are strictly non-accessible during the increm
ental phases. We solve this by training the policy function of RMM on pseudo CIL
 tasks, e.g., the tasks built on the data of the zeroth phase, and then applying
 it to target tasks. RMM propagates two levels of actions: Level-1 determines ho
w to split the memory between old and new classes, and Level-2 allocates memory
for each specific class. In essence, it is an optimizable and general method for
 memory management that can be used in any replaying-based CIL method. For evalu
ation, we plug RMM into two top-performing baselines (LUCIR+AANets and POD+AANet
s [28]) and conduct experiments on three benchmarks (CIFAR-100, ImageNet-Subset,
 and ImageNet-Full). Our results show clear improvements, e.g., boosting POD+AAN
ets by 3.6%, 4.4%, and 1.9% in the 25-Phase settings of the above benchmarks, re
spectively. The code is available at https://class-il.mpi-inf.mpg.de/rmm/.
****
Learning Compact Representations of Neural Networks using DiscriminAtive Masking
 (DAM)
Jie Bu, Arka Daw, M. Maruf, Anuj Karpatne
A central goal in deep learning is to learn compact representations of features
 at every layer of a neural network, which is useful for both unsupervised repre
sentation learning and structured network pruning. While there is a growing bod
y of work in structured pruning, current state-of-the-art methods suffer from t
wo key limitations: (i) instability during training, and (ii) need for an additi
onal step of fine-tuning, which is resource-intensive. At the core of these limi
tations is the lack of a systematic approach that jointly prunes and refines wei
ghts during training in a single stage, and does not require any fine-tuning upo
n convergence to achieve state-of-the-art performance. We present a novel single
-stage structured pruning method termed DiscriminAtive Masking (DAM). The key in
tuition behind DAM is to discriminatively prefer some of the neurons to be refin

ed during the training process, while gradually masking out other neurons. We sh ow that our proposed DAM approach has remarkably good performance over a diverse range of applications in representation learning and structured pruning, includ ing dimensionality reduction, recommendation system, graph representation learni ng, and structured pruning for image classification. We also theoretically show that the learning objective of DAM is directly related to minimizing the L_0 nor m of the masking layer. All of our codes and datasets are available https://gith ub.com/jayroxis/dam-pytorch.
****
Neural Auto-Curricula in Two-Player Zero-Sum Games
Xidong Feng, Oliver Slumbers, Ziyu Wan, Bo Liu, Stephen McAleer, Ying Wen, Jun W ang, Yaodong Yang
When solving two-player zero-sum games, multi-agent reinforcement learning (MARL) algorithms often create populations of agents where, at each iteration, a new agent is discovered as the best response to a mixture over the opponent populati on. Within such a process, the update rules of "who to compete with" (i.e., the opponent mixture) and "how to beat them" (i.e., finding best responses) are unde rpinned by manually developed game theoretical principles such as fictitious pla y and Double Oracle. In this paper, we introduce a novel framework—Neural Auto-C urricula (NAC)—that leverages meta-gradient descent to automate the discovery of the learning update rule without explicit human design. Specifically, we parame terise the opponent selection module by neural networks and the best-response mo dule by optimisation subroutines, and update their parameters solely via interac tion with the game engine, where both players aim to minimise their exploitabili ty. Surprisingly, even without human design, the discovered MARL algorithms achi eve competitive or even better performance with the state-of-the-art population-based game solvers (e.g., PSRO) on Games of Skill, differentiable Lotto, non-tra nsitive Mixture Games, Iterated Matching Pennies, and Kuhn Poker. Additionally, we show that NAC is able to generalise from small games to large games, for exam ple training on Kuhn Poker and outperforming PSRO on Leduc Poker. Our work inspi res a promising future direction to discover general MARL algorithms solely from data.
****
ImageBART: Bidirectional Context with Multinomial Diffusion for Autoregressive I mage Synthesis
Patrick Esser, Robin Rombach, Andreas Blattmann, Bjorn Ommer
Autoregressive models and their sequential factorization of the data likelihood have recently demonstrated great potential for image representation and synthesi s. Nevertheless, they incorporate image context in a linear 1D order by attendin g only to previously synthesized image patches above or to the left. Not only is this unidirectional, sequential bias of attention unnatural for images as it di sregards large parts of a scene until synthesis is almost complete. It also proc esses the entire image on a single scale, thus ignoring more global contextual i nformation up to the gist of the entire scene. As a remedy we incorporate a coar se-to-fine hierarchy of context by combining the autoregressive formulation with a multinomial diffusion process: Whereas a multistage diffusion process success ively compresses and removes information to coarsen an image, we train a Markov chain to invert this process. In each stage, the resulting autoregressive ImageB ART model progressively incorporates context from previous stages in a coarse-to -fine manner. Experiments demonstrate the gain over current autoregressive model s, continuous diffusion probabilistic models, and latent variable models. Moreov er, the approach enables to control the synthesis process and to trade compressi on rate against reconstruction accuracy, while still guaranteeing visually plaus ible results.
****
From global to local MDI variable importances for random forests and when they a re Shapley values
Antonio Sutera, Gilles Louppe, Van Anh Huynh-Thu, Louis Wehenkel, Pierre Geurts
Random forests have been widely used for their ability to provide so-called impo rtance measures, which give insight at a global (per dataset) level on the relev

ance of input variables to predict a certain output. On the other hand, methods based on Shapley values have been introduced to refine the analysis of feature relevance in tree-based models to a local (per instance) level. In this context, we first show that the global Mean Decrease of Impurity (MDI) variable importance scores correspond to Shapley values under some conditions. Then, we derive a local MDI importance measure of variable relevance, which has a very natural connection with the global MDI measure and can be related to a new notion of local feature relevance. We further link local MDI importances with Shapley values and discuss them in the light of related measures from the literature. The measures are illustrated through experiments on several classification and regression problems.
****
Adversarial Robustness of Streaming Algorithms through Importance Sampling
Vladimir Braverman, Avinatan Hassidim, Yossi Matias, Mariano Schain, Sandeep Silwal, Samson 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.
****
Tractable Regularization of Probabilistic Circuits
Anji Liu, Guy Van den Broeck
Probabilistic Circuits (PCs) are a promising avenue for probabilistic modeling. They combine advantages of probabilistic graphical models (PGMs) with those of neural networks (NNs). Crucially, however, they are tractable probabilistic models, supporting efficient and exact computation of many probabilistic inference queries, such as marginals and MAP. Further, since PCs are structured computation graphs, they can take advantage of deep-learning-style parameter updates, which greatly improves their scalability. However, this innovation also makes PCs prone to overfitting, which has been observed in many standard benchmarks. Despite the existence of abundant regularization techniques for both PGMs and NNs, they are not effective enough when applied to PCs. Instead, we re-think regularization for PCs and propose two intuitive techniques, data softening and entropy regularization, that both take advantage of PCs' tractability and still have an efficient implementation as a computation graph. Specifically, data softening provides a principled way to add uncertainty in datasets in closed form, which implicitly regularizes PC parameters. To learn parameters from a softened dataset, PCs only need linear time by virtue of their tractability. In entropy regularization, the exact entropy of the distribution encoded by a PC can be regularized directly, which is again infeasible for most other density estimation models. We show that both methods consistently improve the generalization performance of a wide variety of PCs. Moreover, when paired with a simple PC structure, we achieved state-of-the-art results on 10 out of 20 standard discrete density estimation benchmarks. Open-source code and experiments are available at https://github.com/UCLA-StarAI/Tractable-PC-Regularization.
****
On Interaction Between Augmentations and Corruptions in Natural Corruption Robustness
Eric Mintun, Alexander Kirillov, Saining Xie
Invariance to a broad array of image corruptions, such as warping, noise, or color shifts, is an important aspect of building robust models in computer vision. Recently, several new data augmentations have been proposed that significantly improve performance on ImageNet-C, a benchmark of such corruptions. However, there is still a lack of basic understanding on the relationship between data augmentations and test-time corruptions. To this end, we develop a feature space for image transforms, and then use a new measure in this space between augmentations and corruptions called the Minimal Sample Distance to demonstrate there is a strong correlation between similarity and performance. We then investigate recent data augmentations and observe a significant degradation in corruption robustness when the test-time corruptions are sampled to be perceptually dissimilar from I

mageNet-C in this feature space. Our results suggest that test error can be improved by training on perceptually similar augmentations, and data augmentations may not generalize well beyond the existing benchmark. We hope our results and tools will allow for more robust progress towards improving robustness to image corruptions. We provide code at https://github.com/facebookresearch/augmentation-corruption.
****
Dynamic Distillation Network for Cross-Domain Few-Shot Recognition with Unlabeled Data
Ashraful Islam, Chun-Fu (Richard) Chen, Rameswar Panda, Leonid Karlinsky, Rogerio Feris, Richard J. Radke
Most existing works in few-shot learning rely on meta-learning the network on a large base dataset which is typically from the same domain as the target dataset. We tackle the problem of cross-domain few-shot learning where there is a large shift between the base and target domain. The problem of cross-domain few-shot recognition with unlabeled target data is largely unaddressed in the literature. STARTUP was the first method that tackles this problem using self-training. However, it uses a fixed teacher pretrained on a labeled base dataset to create soft labels for the unlabeled target samples. As the base dataset and unlabeled dataset are from different domains, projecting the target images in the class-domain of the base dataset with a fixed pretrained model might be sub-optimal. We propose a simple dynamic distillation-based approach to facilitate unlabeled images from the novel/base dataset. We impose consistency regularization by calculating predictions from the weakly-augmented versions of the unlabeled images from a teacher network and matching it with the strongly augmented versions of the same images from a student network. The parameters of the teacher network are updated as exponential moving average of the parameters of the student network. We show that the proposed network learns representation that can be easily adapted to the target domain even though it has not been trained with target-specific classes during the pretraining phase. Our model outperforms the current state-of-the art method by 4.4% for 1-shot and 3.6% for 5-shot classification in the BSCD-FSL benchmark, and also shows competitive performance on traditional in-domain few-shot learning task.
****
Hypergraph Propagation and Community Selection for Objects Retrieval
Guoyuan An, Yuchi Huo, Sung-eui Yoon
Spatial verification is a crucial technique for particular object retrieval. It utilizes spatial information for the accurate detection of true positive images. However, existing query expansion and diffusion methods cannot efficiently propagate the spatial information in an ordinary graph with scalar edge weights, resulting in low recall or precision. To tackle these problems, we propose a novel hypergraph-based framework that efficiently propagates spatial information in query time and retrieves an object in the database accurately. Additionally, we propose using the image graph's structure information through community selection technique, to measure the accuracy of the initial search result and to provide correct starting points for hypergraph propagation without heavy spatial verification computations. Experiment results on ROxford and RParis show that our method significantly outperforms the existing query expansion and diffusion methods.
****
Deep learning is adaptive to intrinsic dimensionality of model smoothness in anisotropic Besov space
Taiji Suzuki, Atsushi Nitanda
Deep learning has exhibited superior performance for various tasks, especially for high-dimensional datasets, such as images. To understand this property, we investigate the approximation and estimation ability of deep learning on {\it anisotropic Besov spaces}.The anisotropic Besov space is characterized by direction-dependent smoothness and includes several function classes that have been investigated thus far.We demonstrate that the approximation error and estimation error of deep learning only depend on the average value of the smoothness parameters in all directions. Consequently, the curse of dimensionality can be avoided if t

he smoothness of the target function is highly anisotropic.Unlike existing studies, our analysis does not require a low-dimensional structure of the input data. We also investigate the minimax optimality of deep learning and compare its performance with that of the kernel method (more generally, linear estimators).The results show that deep learning has better dependence on the input dimensionality if the target function possesses anisotropic smoothness, and it achieves an adaptive rate for functions with spatially inhomogeneous smoothness.
****

QuPeD: Quantized Personalization via Distillation with Applications to Federated Learning

Kaan Ozkara, Navjot Singh, Deepesh Data, Suhas Diggavi

Traditionally, federated learning (FL) aims to train a single global model while collaboratively using multiple clients and a server. Two natural challenges that FL algorithms face are heterogeneity in data across clients and collaboration of clients with diverse resources. In this work, we introduce a quantized and personalized FL algorithm QuPeD that facilitates collective (personalized model compression) training via knowledge distillation (KD) among clients who have access to heterogeneous data and resources. For personalization, we allow clients to learn compressed personalized models with different quantization parameters and model dimensions/structures. Towards this, first we propose an algorithm for learning quantized models through a relaxed optimization problem, where quantization values are also optimized over. When each client participating in the (federated) learning process has different requirements of the compressed model (both in model dimension and precision), we formulate a compressed personalization framework by introducing knowledge distillation loss for local client objectives collaborating through a global model. We develop an alternating proximal gradient update for solving this compressed personalization problem, and analyze its convergence properties. Numerically, we validate that QuPeD outperforms competing personalized FL methods, FedAvg, and local training of clients in various heterogeneous settings.
****

Model Adaptation: Historical Contrastive Learning for Unsupervised Domain Adaptation without Source Data

Jiaxing Huang, Dayan Guan, Aoran Xiao, Shijian Lu

Unsupervised domain adaptation aims to align a labeled source domain and an unlabeled target domain, but it requires to access the source data which often raises concerns in data privacy, data portability and data transmission efficiency. We study unsupervised model adaptation (UMA), or called Unsupervised Domain Adaptation without Source Data, an alternative setting that aims to adapt source-trained models towards target distributions without accessing source data. To this end, we design an innovative historical contrastive learning (HCL) technique that exploits historical source hypothesis to make up for the absence of source data in UMA. HCL addresses the UMA challenge from two perspectives. First, it introduces historical contrastive instance discrimination (HCID) that learns from target samples by contrasting their embeddings which are generated by the currently adapted model and the historical models. With the historical models, HCID encourages UMA to learn instance-discriminative target representations while preserving the source hypothesis. Second, it introduces historical contrastive category discrimination (HCCD) that pseudo-labels target samples to learn category-discriminative target representations. Specifically, HCCD re-weights pseudo labels according to their prediction consistency across the current and historical models. Extensive experiments show that HCL outperforms and state-of-the-art methods consistently across a variety of visual tasks and setups.
****

The Out-of-Distribution Problem in Explainability and Search Methods for Feature Importance Explanations

Peter Hase, Harry Xie, Mohit Bansal

Feature importance (FI) estimates are a popular form of explanation, and they are commonly created and evaluated by computing the change in model confidence caused by removing certain input features at test time. For example, in the standar

d Sufficiency metric, only the top-k most important tokens are kept. In this pap
er, we study several under-explored dimensions of FI explanations, providing con
ceptual and empirical improvements for this form of explanation. First, we advan
ce a new argument for why it can be problematic to remove features from an input
 when creating or evaluating explanations: the fact that these counterfactual in
puts are out-of-distribution (OOD) to models implies that the resulting explanat
ions are socially misaligned. The crux of the problem is that the model prior an
d random weight initialization influence the explanations (and explanation metri
cs) in unintended ways. To resolve this issue, we propose a simple alteration to
 the model training process, which results in more socially aligned explanations
 and metrics. Second, we compare among five approaches for removing features fro
m model inputs. We find that some methods produce more OOD counterfactuals than
others, and we make recommendations for selecting a feature-replacement function
. Finally, we introduce four search-based methods for identifying FI explanation
s and compare them to strong baselines, including LIME, Anchors, and Integrated
Gradients. Through experiments with six diverse text classification datasets, we
 find that the only method that consistently outperforms random search is a Para
llel Local Search (PLS) that we introduce. Improvements over the second best met
hod are as large as 5.4 points for Sufficiency and 17 points for Comprehensivene
ss.
****

Control Variates for Slate Off-Policy Evaluation
Nikos Vlassis, Ashok Chandrashekar, Fernando Amat, Nathan Kallus
We study the problem of off-policy evaluation from batched contextual bandit dat
a with multidimensional actions, often termed slates. The problem is common to r
ecommender systems and user-interface optimization, and it is particularly chall
enging because of the combinatorially-sized action space. Swaminathan et al. (20
17) have proposed the pseudoinverse (PI) estimator under the assumption that the
 conditional mean rewards are additive in actions. Using control variates, we co
nsider a large class of unbiased estimators that includes as specific cases the
PI estimator and (asymptotically) its self-normalized variant. By optimizing ove
r this class, we obtain new estimators with risk improvement guarantees over bot
h the PI and the self-normalized PI estimators. Experiments with real-world reco
mmender data as well as synthetic data validate these improvements in practice.
****

Stabilizing Deep Q-Learning with ConvNets and Vision Transformers under Data Aug
mentation
Nicklas Hansen, Hao Su, Xiaolong Wang
While agents trained by Reinforcement Learning (RL) can solve increasingly chall
enging tasks directly from visual observations, generalizing learned skills to n
ovel environments remains very challenging. Extensive use of data augmentation i
s a promising technique for improving generalization in RL, but it is often foun
d to decrease sample efficiency and can even lead to divergence. In this paper,
we investigate causes of instability when using data augmentation in common off-
policy RL algorithms. We identify two problems, both rooted in high-variance Q-t
argets. Based on our findings, we propose a simple yet effective technique for s
tabilizing this class of algorithms under augmentation. We perform extensive emp
irical evaluation of image-based RL using both ConvNets and Vision Transformers
(ViT) on a family of benchmarks based on DeepMind Control Suite, as well as in r
obotic manipulation tasks. Our method greatly improves stability and sample effi
ciency of ConvNets under augmentation, and achieves generalization results compe
titive with state-of-the-art methods for image-based RL in environments with uns
een visuals. We further show that our method scales to RL with ViT-based archite
ctures, and that data augmentation may be especially important in this setting.
****

On Effective Scheduling of Model-based Reinforcement Learning
Hang Lai, Jian Shen, Weinan Zhang, Yimin Huang, Xing Zhang, Ruiming Tang, Yong Y
u, Zhenguo Li
Model-based reinforcement learning has attracted wide attention due to its super
ior sample efficiency. Despite its impressive success so far, it is still unclea

r how to appropriately schedule the important hyperparameters to achieve adequat e performance, such as the real data ratio for policy optimization in Dyna-style model-based algorithms. In this paper, we first theoretically analyze the role of real data in policy training, which suggests that gradually increasing the ra tio of real data yields better performance. Inspired by the analysis, we propose a framework named AutoMBPO to automatically schedule the real data ratio as wel l as other hyperparameters in training model-based policy optimization (MBPO) al gorithm, a representative running case of model-based methods. On several contin uous control tasks, the MBPO instance trained with hyperparameters scheduled by AutoMBPO can significantly surpass the original one, and the real data ratio sch edule found by AutoMBPO shows consistency with our theoretical analysis.
****
Removing Inter-Experimental Variability from Functional Data in Systems Neurosci ence
Dominic Gonschorek, Larissa Höfling, Klaudia P. Szatko, Katrin Franke, Timm Schu bert, Benjamin Dunn, Philipp Berens, David Klindt, Thomas Euler
Integrating data from multiple experiments is common practice in systems neurosc ience but it requires inter-experimental variability to be negligible compared t o the biological signal of interest. This requirement is rarely fulfilled; syste matic changes between experiments can drastically affect the outcome of complex analysis pipelines. Modern machine learning approaches designed to adapt models across multiple data domains offer flexible ways of removing inter-experimental variability where classical statistical methods often fail. While applications o f these methods have been mostly limited to single-cell genomics, in this work, we develop a theoretical framework for domain adaptation in systems neuroscience . We implement this in an adversarial optimization scheme that removes inter-exp erimental variability while preserving the biological signal. We compare our met hod to previous approaches on a large-scale dataset of two-photon imaging record ings of retinal bipolar cell responses to visual stimuli. This dataset provides a unique benchmark as it contains biological signal from well-defined cell types that is obscured by large inter-experimental variability. In a supervised setti ng, we compare the generalization performance of cell type classifiers across ex periments, which we validate with anatomical cell type distributions from electr on microscopy data. In an unsupervised setting, we remove inter-experimental var iability from the data which can then be fed into arbitrary downstream analyses. In both settings, we find that our method achieves the best trade-off between r emoving inter-experimental variability and preserving biological signal. Thus, w e offer a flexible approach to remove inter-experimental variability and integra te datasets across experiments in systems neuroscience. Code available at https: //github.com/eulerlab/rave.
****
Learning Knowledge Graph-based World Models of Textual Environments
Prithviraj Ammanabrolu, Mark Riedl
World models improve a learning agent's ability to efficiently operate in intera ctive and situated environments. This work focuses on the task of building world models of text-based game environments. Text-based games, or interactive narrat ives, are reinforcement learning environments in which agents perceive and inter act with the world using textual natural language. These environments contain lo ng, multi-step puzzles or quests woven through a world that is filled with hundr eds of characters, locations, and objects. Our world model learns to simultaneou sly: (1) predict changes in the world caused by an agent's actions when represen ting the world as a knowledge graph; and (2) generate the set of contextually re levant natural language actions required to operate in the world. We frame this task as a Set of Sequences generation problem by exploiting the inherent structu re of knowledge graphs and actions and introduce both a transformer-based multi- task architecture and a loss function to train it. A zero-shot ablation study on never-before-seen textual worlds shows that our methodology significantly outpe rforms existing textual world modeling techniques as well as the importance of e ach of our contributions.
****

Damped Anderson Mixing for Deep Reinforcement Learning: Acceleration, Convergenc e, and Stabilization
Ke Sun, Yafei Wang, Yi Liu, yingnan zhao, Bo Pan, Shangling Jui, Bei Jiang, Ling long Kong
Anderson mixing has been heuristically applied to reinforcement learning (RL) al gorithms for accelerating convergence and improving the sampling efficiency of d eep RL. Despite its heuristic improvement of convergence, a rigorous mathematica l justification for the benefits of Anderson mixing in RL has not yet been put f orward. In this paper, we provide deeper insights into a class of acceleration s chemes built on Anderson mixing that improve the convergence of deep RL algorith ms. Our main results establish a connection between Anderson mixing and quasi-Ne wton methods and prove that Anderson mixing increases the convergence radius of policy iteration schemes by an extra contraction factor. The key focus of the an alysis roots in the fixed-point iteration nature of RL. We further propose a sta bilization strategy by introducing a stable regularization term in Anderson mixi ng and a differentiable, non-expansive MellowMax operator that can allow both fa ster convergence and more stable behavior. Extensive experiments demonstrate tha t our proposed method enhances the convergence, stability, and performance of RL algorithms.
****
Approximate Decomposable Submodular Function Minimization for Cardinality-Based Components
Nate Veldt, Austin R. Benson, Jon Kleinberg
Minimizing a sum of simple submodular functions of limited support is a special case of general submodular function minimization that has seen numerous applicat ions in machine learning. We develop faster techniques for instances where compo nents in the sum are cardinality-based, meaning they depend only on the size of the input set. This variant is one of the most widely applied in practice, encom passing, e.g., common energy functions arising in image segmentation and recent generalized hypergraph cut functions. We develop the first approximation algorit hms for this problem, where the approximations can be quickly computed via reduc tion to a sparse graph cut problem, with graph sparsity controlled by the desire d approximation factor. Our method relies on a new connection between sparse gra ph reduction techniques and piecewise linear approximations to concave functions . Our sparse reduction technique leads to significant improvements in theoretica l runtimes, as well as substantial practical gains in problems ranging from benc hmark image segmentation tasks to hypergraph clustering problems.
****
Episodic Multi-agent Reinforcement Learning with Curiosity-driven Exploration
Lulu Zheng, Jiarui Chen, Jianhao Wang, Jiamin He, Yujing Hu, Yingfeng Chen, Chan gjie Fan, Yang Gao, Chongjie Zhang
Efficient exploration in deep cooperative multi-agent reinforcement learning (MA RL) still remains challenging in complex coordination problems. In this paper, w e introduce a novel Episodic Multi-agent reinforcement learning with Curiosity-d riven exploration, called EMC. We leverage an insight of popular factorized MARL algorithms that the ``induced" individual Q-values, i.e., the individual utilit y functions used for local execution, are the embeddings of local action-obser vation histories, and can capture the interaction between agents due to reward b ackpropagation during centralized training. Therefore, we use prediction errors of individual Q-values as intrinsic rewards for coordinated exploration and util ize episodic memory to exploit explored informative experience to boost policy t raining. As the dynamics of an agent's individual Q-value function captures the novelty of states and the influence from other agents, our intrinsic reward can induce coordinated exploration to new or promising states. We illustrate the adv antages of our method by didactic examples, and demonstrate its significant outp erformance over state-of-the-art MARL baselines on challenging tasks in the Star Craft II micromanagement benchmark.
****
Two Sides of Meta-Learning Evaluation: In vs. Out of Distribution
Amrith Setlur, Oscar Li, Virginia Smith

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

Debiased Visual Question Answering from Feature and Sample Perspectives
Zhiquan Wen, Guanghui Xu, Mingkui Tan, Qingyao Wu, Qi Wu
Visual question answering (VQA) is designed to examine the visual-textual reasoning ability of an intelligent agent. However, recent observations show that many VQA models may only capture the biases between questions and answers in a dataset rather than showing real reasoning abilities. For example, given a question, some VQA models tend to output the answer that occurs frequently in the dataset and ignore the images. To reduce this tendency, existing methods focus on weakening the language bias. Meanwhile, only a few works also consider vision bias implicitly. However, these methods introduce additional annotations or show unsatisfactory performance. Moreover, not all biases are harmful to the models. Some “biases” learnt from datasets represent natural rules of the world and can help limit the range of answers. Thus, how to filter and remove the true negative biases in language and vision modalities remain a major challenge. In this paper, we propose a method named D-VQA to alleviate the above challenges from the feature and sample perspectives. Specifically, from the feature perspective, we build a question-to-answer and vision-to-answer branch to capture the language and vision biases, respectively. Next, we apply two unimodal bias detection modules to explicitly recognise and remove the negative biases. From the sample perspective, we construct two types of negative samples to assist the training of the models, without introducing additional annotations. Extensive experiments on the VQA-CP v2 and VQA v2 datasets demonstrate the effectiveness of our D-VQA method.
****

Towards a Unified Game-Theoretic View of Adversarial Perturbations and Robustness
Jie Ren, Die Zhang, Yisen Wang, Lu Chen, Zhanpeng Zhou, Yiting Chen, Xu Cheng, Xin Wang, Meng Zhou, Jie Shi, Quanshi Zhang
This paper provides a unified view to explain different adversarial attacks and defense methods, i.e. the view of multi-order interactions between input variables of DNNs. Based on the multi-order interaction, we discover that adversarial attacks mainly affect high-order interactions to fool the DNN. Furthermore, we find that the robustness of adversarially trained DNNs comes from category-specific low-order interactions. Our findings provide a potential method to unify adversarial perturbations and robustness, which can explain the existing robustness-boosting methods in a principle way. Besides, our findings also make a revision of previous inaccurate understanding of the shape bias of adversarially learned features. Our code is available online at https://github.com/Jie-Ren/A-Unified-Game-Theoretic-Interpretation-of-Adversarial-Robustness.
****

On the Out-of-distribution Generalization of Probabilistic Image Modelling
Mingtian Zhang, Andi Zhang, Steven McDonagh
Out-of-distribution (OOD) detection and lossless compression constitute two problems that can be solved by the training of probabilistic models on a first dataset with subsequent likelihood evaluation on a second dataset, where data distributions differ. By defining the generalization of probabilistic models in terms of likelihood we show that, in the case of image models, the OOD generalization ability is dominated by local features. This motivates our proposal of a Local Autoregressive model that exclusively models local image features towards improving OOD performance. We apply the proposed model to OOD detection tasks and achieve state-of-the-art unsupervised OOD detection performance without the introduction of additional data. Additionally, we employ our model to build a new lossless image compressor: NeLLoC (Neural Local Lossless Compressor) and report state-of-the-art compression rates and model size.
****

Exploiting Local Convergence of Quasi-Newton Methods Globally: Adaptive Sample S ize Approach

Qiujiang Jin, Aryan Mokhtari

In this paper, we study the application of quasi-Newton methods for solving empirical risk minimization (ERM) problems defined over a large dataset. Traditional deterministic and stochastic quasi-Newton methods can be executed to solve such problems; however, it is known that their global convergence rate may not be better than first-order methods, and their local superlinear convergence only appears towards the end of the learning process. In this paper, we use an adaptive sample size scheme that exploits the superlinear convergence of quasi-Newton methods globally and throughout the entire learning process. The main idea of the proposed adaptive sample size algorithms is to start with a small subset of data points and solve their corresponding ERM problem within its statistical accuracy, and then enlarge the sample size geometrically and use the optimal solution of the problem corresponding to the smaller set as an initial point for solving the subsequent ERM problem with more samples. We show that if the initial sample size is sufficiently large and we use quasi-Newton methods to solve each subproblem, the subproblems can be solved superlinearly fast (after at most three iterations), as we guarantee that the iterates always stay within a neighborhood that quasi-Newton methods converge superlinearly. Numerical experiments on various datasets confirm our theoretical results and demonstrate the computational advantages of our method.

****

PDE-GCN: Novel Architectures for Graph Neural Networks Motivated by Partial Differential Equations

Moshe Eliasof, Eldad Haber, Eran Treister

Graph neural networks are increasingly becoming the go-to approach in various fields such as computer vision, computational biology and chemistry, where data are naturally explained by graphs. However, unlike traditional convolutional neural networks, deep graph networks do not necessarily yield better performance than shallow graph networks. This behavior usually stems from the over-smoothing phenomenon. In this work, we propose a family of architecturesto control this behavior by design. Our networks are motivated by numerical methods for solving Partial Differential Equations (PDEs) on manifolds, and as such, their behavior can be explained by similar analysis. Moreover, as we demonstrate using an extensive set of experiments, our PDE-motivated networks can generalize and be effective for various types of problems from different fields. Our architectures obtain better or on par with the current state-of-the-art results for problems that are typically approached using different architectures.

****

Information Directed Reward Learning for Reinforcement Learning

David Lindner, Matteo Turchetta, Sebastian Tschiatschek, Kamil Ciosek, Andreas Krause

For many reinforcement learning (RL) applications, specifying a reward is difficult. In this paper, we consider an RL setting where the agent can obtain information about the reward only by querying an expert that can, for example, evaluate individual states or provide binary preferences over trajectories. From such expensive feedback, we aim to learn a model of the reward function that allows standard RL algorithms to achieve high expected return with as few expert queries as possible. For this purpose, we propose Information Directed Reward Learning (IDRL), which uses a Bayesian model of the reward function and selects queries that maximize the information gain about the difference in return between potentially optimal policies. In contrast to prior active reward learning methods designed for specific types of queries, IDRL naturally accommodates different query types. Moreover, by shifting the focus from reducing the reward approximation error to improving the policy induced by the reward model, it achieves similar or better performance with significantly fewer queries. We support our findings with extensive evaluations in multiple environments and with different types of queries.

****

SSMF: Shifting Seasonal Matrix Factorization

Koki Kawabata, Siddharth Bhatia, Rui Liu, Mohit Wadhwa, Bryan Hooi
Given taxi-ride counts information between departure and destination locations, how can we forecast their future demands? In general, given a data stream of events with seasonal patterns that innovate over time, how can we effectively and efficiently forecast future events? In this paper, we propose Shifting Seasonal Matrix Factorization approach, namely SSMF, that can adaptively learn multiple seasonal patterns (called regimes), as well as switching between them. Our proposed method has the following properties: (a) it accurately forecasts future events by detecting regime shifts in seasonal patterns as the data stream evolves; (b) it works in an online setting, i.e., processes each observation in constant time and memory; (c) it effectively realizes regime shifts without human intervention by using a lossless data compression scheme. We demonstrate that our algorithm outperforms state-of-the-art baseline methods by accurately forecasting upcoming events on three real-world data streams.
****
Associative Memories via Predictive Coding
Tommaso Salvatori, Yuhang Song, Yujian Hong, Lei Sha, Simon Frieder, Zhenghua Xu, Rafal Bogacz, Thomas Lukasiewicz
Associative memories in the brain receive and store patterns of activity registered by the sensory neurons, and are able to retrieve them when necessary. Due to their importance in human intelligence, computational models of associative memories have been developed for several decades now. In this paper, we present a novel neural model for realizing associative memories, which is based on a hierarchical generative network that receives external stimuli via sensory neurons. It is trained using predictive coding, an error-based learning algorithm inspired by information processing in the cortex. To test the model's capabilities, we perform multiple retrieval experiments from both corrupted and incomplete data points. In an extensive comparison, we show that this new model outperforms in retrieval accuracy and robustness popular associative memory models, such as autoencoders trained via backpropagation, and modern Hopfield networks. In particular, in completing partial data points, our model achieves remarkable results on natural image datasets, such as ImageNet, with a surprisingly high accuracy, even when only a tiny fraction of pixels of the original images is presented. Our model provides a plausible framework to study learning and retrieval of memories in the brain, as it closely mimics the behavior of the hippocampus as a memory index and generative model.
****
Robust and differentially private mean estimation
Xiyang Liu, Weihao Kong, Sham Kakade, Sewoong Oh
In statistical learning and analysis from shared data, which is increasingly widely adopted in platforms such as federated learning and meta-learning, there are two major concerns: privacy and robustness. Each participating individual should be able to contribute without the fear of leaking one's sensitive information. At the same time, the system should be robust in the presence of malicious participants inserting corrupted data. Recent algorithmic advances in learning from shared data focus on either one of these threats, leaving the system vulnerable to the other. We bridge this gap for the canonical problem of estimating the mean from i.i.d.~samples. We introduce PRIME, which is the first efficient algorithm that achieves both privacy and robustness for a wide range of distributions. We further complement this result with a novel exponential time algorithm that improves the sample complexity of PRIME, achieving a near-optimal guarantee and matching that of a known lower bound for (non-robust) private mean estimation. This proves that there is no extra statistical cost to simultaneously guaranteeing privacy and robustness.
****
Adaptable Agent Populations via a Generative Model of Policies
Kenneth Derek, Phillip Isola
In the natural world, life has found innumerable ways to survive and often thrive. Between and even within species, each individual is in some manner unique, and this diversity lends adaptability and robustness to life. In this work, we aim

 to learn a space of diverse and high-reward policies in a given environment. To this end, we introduce a generative model of policies for reinforcement learning, which maps a low-dimensional latent space to an agent policy space. Our method enables learning an entire population of agent policies, without requiring the use of separate policy parameters. Just as real world populations can adapt and evolve via natural selection, our method is able to adapt to changes in our environment solely by selecting for policies in latent space. We test our generative model's capabilities in a variety of environments, including an open-ended grid-world and a two-player soccer environment. Code, visualizations, and additional experiments can be found at https://kennyderek.github.io/adap/.
****
A No-go Theorem for Robust Acceleration in the Hyperbolic Plane
Linus Hamilton, Ankur Moitra
In recent years there has been significant effort to adapt the key tools and ideas in convex optimization to the Riemannian setting. One key challenge has remained: Is there a Nesterov-like accelerated gradient method for geodesically convex functions on a Riemannian manifold? Recent work has given partial answers and the hope was that this ought to be possible. Here we prove that in a noisy setting, there is no analogue of accelerated gradient descent for geodesically convex functions on the hyperbolic plane. Our results apply even when the noise is exponentially small. The key intuition behind our proof is short and simple: In negatively curved spaces, the volume of a ball grows so fast that information about the past gradients is not useful in the future.
****
Privately Learning Mixtures of Axis-Aligned Gaussians
Ishaq Aden-Ali, Hassan Ashtiani, Christopher Liaw
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 Self-Dissimilarities as Powerful Visual Fingerprints
Idan Kligvasser, Tamar Shaham, Yuval Bahat, Tomer Michaeli
Features extracted from deep layers of classification networks are widely used as image descriptors. Here, we exploit an unexplored property of these features: their internal dissimilarity. While small image patches are known to have similar statistics across image scales, it turns out that the internal distribution of deep features varies distinctively between scales. We show how this deep self dissimilarity (DSD) property can be used as a powerful visual fingerprint. Particularly, we illustrate that full-reference and no-reference image quality measures derived from DSD are highly correlated with human preference. In addition, incorporating DSD as a loss function in training of image restoration networks, leads to results that are at least as photo-realistic as those obtained by GAN based methods, while not requiring adversarial training.
****
Invariant Causal Imitation Learning for Generalizable Policies
Ioana Bica, Daniel Jarrett, Mihaela van der Schaar
Consider learning an imitation policy on the basis of demonstrated behavior from multiple environments, with an eye towards deployment in an unseen environment. Since the observable features from each setting may be different, directly learning individual policies as mappings from features to actions is prone to spurious correlations---and may not generalize well. However, the expert's policy is often a function of a shared latent structure underlying those observable features that is invariant across settings. By leveraging data from multiple environments, we propose Invariant Causal Imitation Learning (ICIL), a novel technique in which we learn a feature representation that is invariant across domains, on the basis of which we learn an imitation policy that matches expert behavior. To cope with transition dynamics mismatch, ICIL learns a shared representation of causal features (for all training environments), that is disentangled from the specific representations of noise variables (for each of those environments). Moreov

er, to ensure that the learned policy matches the observation distribution of the expert's policy, ICIL estimates the energy of the expert's observations and uses a regularization term that minimizes the imitator policy's next state energy. Experimentally, we compare our methods against several benchmarks in control and healthcare tasks and show its effectiveness in learning imitation policies capable of generalizing to unseen environments.
****
CoAtNet: Marrying Convolution and Attention for All Data Sizes
Zihang Dai, Hanxiao Liu, Quoc V Le, Mingxing Tan
Transformers have attracted increasing interests in computer vision, but they still fall behind state-of-the-art convolutional networks. In this work, we show that while Transformers tend to have larger model capacity, their generalization can be worse than convolutional networks due to the lack of the right inductive bias. To effectively combine the strengths from both architectures, we present CoAtNets(pronounced "coat" nets), a family of hybrid models built from two key insights: (1) depthwise Convolution and self-Attention can be naturally unified via a simple relative attention; (2) vertically stacking convolution layers and attention layers in a principled way is surprisingly effective in improving generalization, capacity and efficiency. Experiments show that our CoAtNets achieve state-of-the-art performance under different resource constraints across various datasets: Without extra data, CoAtNet achieves 86.0% ImageNet top-1 accuracy; When pre-trained with 13M images from ImageNet-21K, our CoAtNet achieves 88.56% top-1 accuracy, matching ViT-huge pre-trained with 300M images from JFT-300M while using 23x less data; Notably, when we further scale up CoAtNet with JFT-3B, it achieves 90.88% top-1 accuracy on ImageNet, establishing a new state-of-the-art result.
****
Mixed Supervised Object Detection by Transferring Mask Prior and Semantic Similarity
Yan Liu, Zhijie Zhang, Li Niu, Junjie Chen, Liqing Zhang
Object detection has achieved promising success, but requires large-scale fully-annotated data, which is time-consuming and labor-extensive. Therefore, we consider object detection with mixed supervision, which learns novel object categories using weak annotations with the help of full annotations of existing base object categories. Previous works using mixed supervision mainly learn the class-agnostic objectness from fully-annotated categories, which can be transferred to upgrade the weak annotations to pseudo full annotations for novel categories. In this paper, we further transfer mask prior and semantic similarity to bridge the gap between novel categories and base categories. Specifically, the ability of using mask prior to help detect objects is learned from base categories and transferred to novel categories. Moreover, the semantic similarity between objects learned from base categories is transferred to denoise the pseudo full annotations for novel categories. Experimental results on three benchmark datasets demonstrate the effectiveness of our method over existing methods. Codes are available at https://github.com/bcmi/TraMaS-Weak-Shot-Object-Detection.
****
Celebrating Diversity in Shared Multi-Agent Reinforcement Learning
Chenghao Li, Tonghan Wang, Chengjie Wu, Qianchuan Zhao, Jun Yang, Chongjie Zhang
Recently, deep multi-agent reinforcement learning (MARL) has shown the promise to solve complex cooperative tasks. Its success is partly because of parameter sharing among agents. However, such sharing may lead agents to behave similarly and limit their coordination capacity. In this paper, we aim to introduce diversity in both optimization and representation of shared multi-agent reinforcement learning. Specifically, we propose an information-theoretical regularization to maximize the mutual information between agents' identities and their trajectories, encouraging extensive exploration and diverse individualized behaviors. In representation, we incorporate agent-specific modules in the shared neural network architecture, which are regularized by L1-norm to promote learning sharing among agents while keeping necessary diversity. Empirical results show that our method achieves state-of-the-art performance on Google Research Football and super har

d StarCraft II micromanagement tasks.
****
Rebounding Bandits for Modeling Satiation Effects
Liu Leqi, Fatma Kilinc Karzan, Zachary Lipton, Alan Montgomery
Psychological research shows that enjoyment of many goods is subject to satiatio
n, with short-term satisfaction declining after repeated exposures to the same i
tem. Nevertheless, proposed algorithms for powering recommender systems seldom m
odel these dynamics, instead proceeding as though user preferences were fixed in
 time. In this work, we introduce rebounding bandits, a multi-armed bandit setup
, where satiation dynamics are modeled as time-invariant linear dynamical system
s. Expected rewards for each arm decline monotonically with consecutive exposure
s and rebound towards the initial reward whenever that arm is not pulled. Unlike
 classical bandit algorithms, methods for tackling rebounding bandits must plan
ahead and model-based methods rely on estimating the parameters of the satiation
 dynamics. We characterize the planning problem, showing that the greedy policy
is optimal when the arms exhibit identical deterministic dynamics. To address st
ochastic satiation dynamics with unknown parameters, we propose Explore-Estimate
-Plan, an algorithm that pulls arms methodically, estimates the system dynamics,
 and then plans accordingly.
****
Sample Complexity of Tree Search Configuration: Cutting Planes and Beyond
Maria-Florina F. Balcan, Siddharth Prasad, Tuomas Sandholm, Ellen Vitercik
Cutting-plane methods have enabled remarkable successes in integer programming o
ver the last few decades. State-of-the-art solvers integrate a myriad of cutting
-plane techniques to speed up the underlying tree-search algorithm used to find
optimal solutions. In this paper we provide sample complexity bounds for cut-sel
ection in branch-and-cut (B&C). Given a training set of integer programs sampled
 from an application-specific input distribution and a family of cut selection p
olicies, these guarantees bound the number of samples sufficient to ensure that
using any policy in the family, the size of the tree B&C builds on average over
the training set is close to the expected size of the tree B&C builds. We first
bound the sample complexity of learning cutting planes from the canonical family
 of Chvátal-Gomory cuts. Our bounds handle any number of waves of any number of
cuts and are fine tuned to the magnitudes of the constraint coefficients. Next,
we prove sample complexity bounds for more sophisticated cut selection policies
that use a combination of scoring rules to choose from a family of cuts. Finally
, beyond the realm of cutting planes for integer programming, we develop a gener
al abstraction of tree search that captures key components such as node selectio
n and variable selection. For this abstraction, we bound the sample complexity o
f learning a good policy for building the search tree.
****
IQ-Learn: Inverse soft-Q Learning for Imitation
Divyansh Garg, Shuvam Chakraborty, Chris Cundy, Jiaming Song, Stefano Ermon
In many sequential decision-making problems (e.g., robotics control, game playin
g, sequential prediction), human or expert data is available containing useful i
nformation about the task. However, imitation learning (IL) from a small amount
of expert data can be challenging in high-dimensional environments with complex
dynamics. Behavioral cloning is a simple method that is widely used due to its s
implicity of implementation and stable convergence but doesn't utilize any infor
mation involving the environment’s dynamics. Many existing methods that exploit
dynamics information are difficult to train in practice due to an adversarial op
timization process over reward and policy approximators or biased, high variance
 gradient estimators. We introduce a method for dynamics-aware IL which avoids a
dversarial training by learning a single Q-function, implicitly representing bot
h reward and policy. On standard benchmarks, the implicitly learned rewards show
 a high positive correlation with the ground-truth rewards, illustrating our met
hod can also be used for inverse reinforcement learning (IRL). Our method, Inver
se soft-Q learning (IQ-Learn) obtains state-of-the-art results in offline and on
line imitation learning settings, significantly outperforming existing methods b
oth in the number of required environment interactions and scalability in high-d

imensional spaces, often by more than 3x.
****
Task-Agnostic Undesirable Feature Deactivation Using Out-of-Distribution Data
Dongmin Park, Hwanjun Song, Minseok Kim, Jae-Gil Lee
A deep neural network (DNN) has achieved great success in many machine learning tasks by virtue of its high expressive power. However, its prediction can be easily biased to undesirable features, which are not essential for solving the target task and are even imperceptible to a human, thereby resulting in poor generalization. Leveraging plenty of undesirable features in out-of-distribution (OOD) examples has emerged as a potential solution for de-biasing such features, and a recent study shows that softmax-level calibration of OOD examples can successfully remove the contribution of undesirable features to the last fully-connected layer of a classifier. However, its applicability is confined to the classification task, and its impact on a DNN feature extractor is not properly investigated. In this paper, we propose Taufe, a novel regularizer that deactivates many undesirable features using OOD examples in the feature extraction layer and thus removes the dependency on the task-specific softmax layer. To show the task-agnostic nature of Taufe, we rigorously validate its performance on three tasks, classification, regression, and a mix of them, on CIFAR-10, CIFAR-100, ImageNet, CUB200, and CAR datasets. The results demonstrate that Taufe consistently outperforms the state-of-the-art method as well as the baselines without regularization.
****
Private Non-smooth ERM and SCO in Subquadratic Steps
Janardhan Kulkarni, Yin Tat Lee, Daogao 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.
****
Towards Instance-Optimal Offline Reinforcement Learning with Pessimism
Ming Yin, Yu-Xiang Wang
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Speedy Performance Estimation for Neural Architecture Search
Robin Ru, Clare Lyle, Lisa Schut, Miroslav Fil, Mark van der Wilk, Yarin Gal
Reliable yet efficient evaluation of generalisation performance of a proposed architecture is crucial to the success of neural architecture search (NAS). Traditional approaches face a variety of limitations: training each architecture to completion is prohibitively expensive, early stopped validation accuracy may correlate poorly with fully trained performance, and model-based estimators require large training sets. We instead propose to estimate the final test performance based on a simple measure of training speed. Our estimator is theoretically motivated by the connection between generalisation and training speed, and is also inspired by the reformulation of a PAC-Bayes bound under the Bayesian setting. Our model-free estimator is simple, efficient, and cheap to implement, and does not require hyperparameter-tuning or surrogate training before deployment. We demonstrate on various NAS search spaces that our estimator consistently outperforms other alternatives in achieving better correlation with the true test performance rankings. We further show that our estimator can be easily incorporated into both query-based and one-shot NAS methods to improve the speed or quality of the search.
****
How Tight Can PAC-Bayes be in the Small Data Regime?
Andrew Foong, Wessel Bruinsma, David Burt, 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-auth

ors prior to requesting a name change in the electronic proceedings.
****
Deep Synoptic Monte-Carlo Planning in Reconnaissance Blind Chess
Gregory Clark
This paper introduces deep synoptic Monte Carlo planning (DSMCP) for large imper fect information games. The algorithm constructs a belief state with an unweight ed particle filter and plans via playouts that start at samples drawn from the b elief state. The algorithm accounts for uncertainty by performing inference on " synopses," a novel stochastic abstraction of information states. DSMCP is the ba sis of the program Penumbra, which won the official 2020 reconnaissance blind ch ess competition versus 33 other programs. This paper also evaluates algorithm va riants that incorporate caution, paranoia, and a novel bandit algorithm. Further more, it audits the synopsis features used in Penumbra with per-bit saliency sta tistics.
****
Dynamic Analysis of Higher-Order Coordination in Neuronal Assemblies via De-Spar sified Orthogonal Matching Pursuit
Shoutik Mukherjee, Behtash Babadi
Coordinated ensemble spiking activity is widely observable in neural recordings and central in the study of population codes, with hypothesized roles including robust stimulus representation, interareal communication of neural information, and learning and memory formation. Model-free measures of synchrony characterize the coherence of pairwise activity, but not higher-order interactions; this lim itation is transcended by statistical models of ensemble spiking activity. Howev er, existing model-based analyses often impose assumptions about the relevance o f higher-order interactions and require multiple repeated trials in order to cha racterize dynamics in the correlational structure of ensemble activity. To addre ss these shortcomings, we propose an adaptive greedy filtering algorithm based o n a discretized mark point-process model of ensemble spiking and a corresponding precise statistical inference framework to identify significant coordinated hig her-order spiking activity. In the course of developing the statistical inferenc e procedures, we also show that confidence intervals can be constructed for gree dily estimated parameters. We demonstrate the utility of our proposed methods on simulated neuronal assemblies. Applied to multi-electrode recordings of human c ortical ensembles, our proposed methods provide new insights into the dynamics u nderlying localized population activity during transitions between brain states.
****
Efficient Training of Retrieval Models using Negative Cache
Erik Lindgren, Sashank Reddi, Ruiqi Guo, Sanjiv Kumar
Factorized models, such as two tower neural network models, are widely used for scoring (query, document) pairs in information retrieval tasks. These models are typically trained by optimizing the model parameters to score relevant positive " pairs higher than the irrelevantnegative" ones. While a large set of negatives typically improves the model performance, limited computation and memory budget s place constraints on the number of negatives used during training. In this pap er, we develop a novel negative sampling technique for accelerating training wit h softmax cross-entropy loss. By using cached (possibly stale) item embeddings, our technique enables training with a large pool of negatives with reduced memor y and computation. We also develop a streaming variant of our algorithm geared t owards very large datasets. Furthermore, we establish a theoretical basis for ou r approach by showing that updating a very small fraction of the cache at each i teration can still ensure fast convergence. Finally, we experimentally validate our approach and show that it is efficient and compares favorably with more comp lex, state-of-the-art approaches.
****
Understanding Partial Multi-Label Learning via Mutual Information
Xiuwen Gong, Dong Yuan, Wei Bao
To deal with ambiguities in partial multilabel learning (PML), state-of-the-art methods perform disambiguation by identifying ground-truth labels directly. Howe ver, there is an essential question:“Can the ground-truth labels be identified p

recisely?". If yes, “How can the ground-truth labels be found?". This paper provides affirmative answers to these questions. Instead of adopting hand-made heuristic strategy, we propose a novel Mutual Information Label Identification for Partial Multilabel Learning (MILI-PML), which is derived from a clear probabilistic formulation and could be easily interpreted theoretically from the mutual information perspective, as well as naturally incorporates the feature/label relevancy considerations. Extensive experiments on synthetic and real-world datasets clearly demonstrate the superiorities of the proposed MILI-PML.
****
Environment Generation for Zero-Shot Compositional Reinforcement Learning
Izzeddin Gur, Natasha Jaques, Yingjie Miao, Jongwook Choi, Manoj Tiwari, Honglak Lee, Aleksandra Faust
Many real-world problems are compositional – solving them requires completing interdependent sub-tasks, either in series or in parallel, that can be represented as a dependency graph. Deep reinforcement learning (RL) agents often struggle to learn such complex tasks due to the long time horizons and sparse rewards. To address this problem, we present Compositional Design of Environments (CoDE), which trains a Generator agent to automatically build a series of compositional tasks tailored to the RL agent’s current skill level. This automatic curriculum not only enables the agent to learn more complex tasks than it could have otherwise, but also selects tasks where the agent’s performance is weak, enhancing its robustness and ability to generalize zero-shot to unseen tasks at test-time. We analyze why current environment generation techniques are insufficient for the problem of generating compositional tasks, and propose a new algorithm that addresses these issues. Our results assess learning and generalization across multiple compositional tasks, including the real-world problem of learning to navigate and interact with web pages. We learn to generate environments composed of multiple pages or rooms, and train RL agents capable of completing wide-range of complex tasks in those environments. We contribute two new benchmark frameworks for generating compositional tasks, compositional MiniGrid and gMiniWoB for web navigation. CoDE yields 4x higher success rate than the strongest baseline, and demonstrates strong performance of real websites learned on 3500 primitive tasks.
****
Optimizing Conditional Value-At-Risk of Black-Box Functions
Quoc Phong Nguyen, Zhongxiang Dai, Bryan Kian Hsiang Low, Patrick Jaillet
This paper presents two Bayesian optimization (BO) algorithms with theoretical performance guarantee to maximize the conditional value-at-risk (CVaR) of a black-box function: CV-UCB and CV-TS which are based on the well-established principle of optimism in the face of uncertainty and Thompson sampling, respectively. To achieve this, we develop an upper confidence bound of CVaR and prove the no-regret guarantee of CV-UCB by utilizing an interesting connection between CVaR and value-at-risk (VaR). For CV-TS, though it is straightforwardly performed with Thompson sampling, bounding its Bayesian regret is non-trivial because it requires a tail expectation bound for the distribution of CVaR of a black-box function, which has not been shown in the literature. The performances of both CV-UCB and CV-TS are empirically evaluated in optimizing CVaR of synthetic benchmark functions and simulated real-world optimization problems.
****
E(n) Equivariant Normalizing Flows
Victor Garcia Satorras, Emiel Hoogeboom, Fabian Fuchs, Ingmar Posner, Max Welling
This paper introduces a generative model equivariant to Euclidean symmetries: E(n) Equivariant Normalizing Flows (E-NFs). To construct E-NFs, we take the discriminative E(n) graph neural networks and integrate them as a differential equation to obtain an invertible equivariant function: a continuous-time normalizing flow. We demonstrate that E-NFs considerably outperform baselines and existing methods from the literature on particle systems such as DW4 and LJ13, and on molecules from QM9 in terms of log-likelihood. To the best of our knowledge, this is the first flow that jointly generates molecule features and positions in 3D.
****

Revitalizing CNN Attention via Transformers in Self-Supervised Visual Representation Learning
Chongjian GE, Youwei Liang, YIBING SONG, Jianbo Jiao, Jue Wang, Ping Luo
Studies on self-supervised visual representation learning (SSL) improve encoder backbones to discriminate training samples without labels. While CNN encoders via SSL achieve comparable recognition performance to those via supervised learning, their network attention is under-explored for further improvement. Motivated by the transformers that explore visual attention effectively in recognition scenarios, we propose a CNN Attention REvitalization (CARE) framework to train attentive CNN encoders guided by transformers in SSL. The proposed CARE framework consists of a CNN stream (C-stream) and a transformer stream (T-stream), where each stream contains two branches. C-stream follows an existing SSL framework with two CNN encoders, two projectors, and a predictor. T-stream contains two transformers, two projectors, and a predictor. T-stream connects to CNN encoders and is in parallel to the remaining C-Stream. During training, we perform SSL in both streams simultaneously and use the T-stream output to supervise C-stream. The features from CNN encoders are modulated in T-stream for visual attention enhancement and become suitable for the SSL scenario. We use these modulated features to supervise C-stream for learning attentive CNN encoders. To this end, we revitalize CNN attention by using transformers as guidance. Experiments on several standard visual recognition benchmarks, including image classification, object detection, and semantic segmentation, show that the proposed CARE framework improves CNN encoder backbones to the state-of-the-art performance.
****
A Critical Look at the Consistency of Causal Estimation with Deep Latent Variable Models
Severi Rissanen, Pekka Marttinen
Using deep latent variable models in causal inference has attracted considerable interest recently, but an essential open question is their ability to yield consistent causal estimates. While they have demonstrated promising results and theory exists on some simple model formulations, we also know that causal effects are not even identifiable in general with latent variables. We investigate this gap between theory and empirical results with analytical considerations and extensive experiments under multiple synthetic and real-world data sets, using the causal effect variational autoencoder (CEVAE) as a case study. While CEVAE seems to work reliably under some simple scenarios, it does not estimate the causal effect correctly with a misspecified latent variable or a complex data distribution, as opposed to its original motivation. Hence, our results show that more attention should be paid to ensuring the correctness of causal estimates with deep latent variable models.
****
Improving Robustness using Generated Data
Sven Gowal, Sylvestre-Alvise Rebuffi, Olivia Wiles, Florian Stimberg, Dan Andrei Calian, Timothy A Mann
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
An Analysis of Constant Step Size SGD in the Non-convex Regime: Asymptotic Normality and Bias
Lu Yu, Krishnakumar Balasubramanian, Stanislav Volgushev, Murat A. Erdogdu
Structured non-convex learning problems, for which critical points have favorable statistical properties, arise frequently in statistical machine learning. Algorithmic convergence and statistical estimation rates are well-understood for such problems. However, quantifying the uncertainty associated with the underlying training algorithm is not well-studied in the non-convex setting. In order to address this shortcoming, in this work, we establish an asymptotic normality result for the constant step size stochastic gradient descent (SGD) algorithm---a widely used algorithm in practice. Specifically, based on the relationship between

SGD and Markov Chains [DDB19], we show that the average of SGD iterates is asymptotically normally distributed around the expected value of their unique invariant distribution, as long as the non-convex and non-smooth objective function satisfies a dissipativity property. We also characterize the bias between this expected value and the critical points of the objective function under various local regularity conditions. Together, the above two results could be leveraged to construct confidence intervals for non-convex problems that are trained using the SGD algorithm.
****
Learning to Learn Graph Topologies
Xingyue Pu, Tianyue Cao, Xiaoyun Zhang, Xiaowen Dong, Siheng 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.
****
Invertible Tabular GANs: Killing Two Birds with One Stone for Tabular Data Synthesis
JAEHOON LEE, Jihyeon Hyeong, Jinsung Jeon, Noseong Park, Jihoon Cho
Tabular data synthesis has received wide attention in the literature. This is because available data is often limited, incomplete, or cannot be obtained easily, and data privacy is becoming increasingly important. In this work, we present a generalized GAN framework for tabular synthesis, which combines the adversarial training of GANs and the negative log-density regularization of invertible neural networks. The proposed framework can be used for two distinctive objectives. First, we can further improve the synthesis quality, by decreasing the negative log-density of real records in the process of adversarial training. On the other hand, by increasing the negative log-density of real records, realistic fake records can be synthesized in a way that they are not too much close to real records and reduce the chance of potential information leakage. We conduct experiments with real-world datasets for classification, regression, and privacy attacks. In general, the proposed method demonstrates the best synthesis quality (in terms of task-oriented evaluation metrics, e.g., F1) when decreasing the negative log-density during the adversarial training. If increasing the negative log-density, our experimental results show that the distance between real and fake records increases, enhancing robustness against privacy attacks.
****
Reducing Collision Checking for Sampling-Based Motion Planning Using Graph Neural Networks
Chenning Yu, Sicun Gao
Sampling-based motion planning is a popular approach in robotics for finding paths in continuous configuration spaces. Checking collision with obstacles is the major computational bottleneck in this process. We propose new learning-based methods for reducing collision checking to accelerate motion planning by training graph neural networks (GNNs) that perform path exploration and path smoothing. Given random geometric graphs (RGGs) generated from batch sampling, the path exploration component iteratively predicts collision-free edges to prioritize their exploration. The path smoothing component then optimizes paths obtained from the exploration stage. The methods benefit from the ability of GNNs of capturing geometric patterns from RGGs through batch sampling and generalize better to unseen environments. Experimental results show that the learned components can significantly reduce collision checking and improve overall planning efficiency in challenging high-dimensional motion planning tasks.
****
Sample Complexity Bounds for Active Ranking from Multi-wise Comparisons
Wenbo Ren, Jia Liu, Ness Shroff
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****
Efficient Bayesian network structure learning via local Markov boundary search
Ming Gao, Bryon Aragam
We analyze the complexity of learning directed acyclic graphical models from observational data in general settings without specific distributional assumptions. Our approach is information-theoretic and uses a local Markov boundary search procedure in order to recursively construct ancestral sets in the underlying graphical model. Perhaps surprisingly, we show that for certain graph ensembles, a simple forward greedy search algorithm (i.e. without a backward pruning phase) suffices to learn the Markov boundary of each node. This substantially improves the sample complexity, which we show is at most polynomial in the number of nodes. This is then applied to learn the entire graph under a novel identifiability condition that generalizes existing conditions from the literature. As a matter of independent interest, we establish finite-sample guarantees for the problem of recovering Markov boundaries from data. Moreover, we apply our results to the special case of polytrees, for which the assumptions simplify, and provide explicit conditions under which polytrees are identifiable and learnable in polynomial time. We further illustrate the performance of the algorithm, which is easy to implement, in a simulation study. Our approach is general, works for discrete or continuous distributions without distributional assumptions, and as such sheds light on the minimal assumptions required to efficiently learn the structure of directed graphical models from data.
****
Learning Dynamic Graph Representation of Brain Connectome with Spatio-Temporal Attention
Byung-Hoon Kim, Jong Chul Ye, Jae-Jin Kim
Functional connectivity (FC) between regions of the brain can be assessed by the degree of temporal correlation measured with functional neuroimaging modalities. Based on the fact that these connectivities build a network, graph-based approaches for analyzing the brain connectome have provided insights into the functions of the human brain. The development of graph neural networks (GNNs) capable of learning representation from graph structured data has led to increased interest in learning the graph representation of the brain connectome. Although recent attempts to apply GNN to the FC network have shown promising results, there is still a common limitation that they usually do not incorporate the dynamic characteristics of the FC network which fluctuates over time. In addition, a few studies that have attempted to use dynamic FC as an input for the GNN reported a reduction in performance compared to static FC methods, and did not provide temporal explainability. Here, we propose STAGIN, a method for learning dynamic graph representation of the brain connectome with spatio-temporal attention. Specifically, a temporal sequence of brain graphs is input to the STAGIN to obtain the dynamic graph representation, while novel READOUT functions and the Transformer encoder provide spatial and temporal explainability with attention, respectively. Experiments on the HCP-Rest and the HCP-Task datasets demonstrate exceptional performance of our proposed method. Analysis of the spatio-temporal attention also provide concurrent interpretation with the neuroscientific knowledge, which further validates our method. Code is available at https://github.com/egyptdj/stagin
****
Understanding the Generalization Benefit of Model Invariance from a Data Perspective
Sicheng Zhu, Bang An, Furong Huang
Machine learning models that are developed to be invariant under certain types of data transformations have shown improved generalization in practice. However, a principled understanding of why invariance benefits generalization is limited. Given a dataset, there is often no principled way to select "suitable" data transformations under which model invariance guarantees better generalization. This paper studies the generalization benefit of model invariance by introducing the sample cover induced by transformations, i.e., a representative subset of a dataset that can approximately recover the whole dataset using transformations. For any data transformations, we provide refined generalization bounds for invarian

t models based on the sample cover. We also characterize the "suitability" of a set of data transformations by the sample covering number induced by transformations, i.e., the smallest size of its induced sample covers. We show that we may tighten the generalization bounds for "suitable" transformations that have a small sample covering number. In addition, our proposed sample covering number can be empirically evaluated and thus provides a guidance for selecting transformations to develop model invariance for better generalization. In experiments on multiple datasets, we evaluate sample covering numbers for some commonly used transformations and show that the smaller sample covering number for a set of transformations (e.g., the 3D-view transformation) indicates a smaller gap between the test and training error for invariant models, which verifies our propositions.
****

Improved Variance-Aware Confidence Sets for Linear Bandits and Linear Mixture MDP

Zihan Zhang, Jiaqi Yang, Xiangyang Ji, Simon S. Du

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 Should Pre-Trained Language Models Be Fine-Tuned Towards Adversarial Robustness?

Xinshuai Dong, Anh Tuan Luu, Min Lin, Shuicheng Yan, Hanwang Zhang

The fine-tuning of pre-trained language models has a great success in many NLP fields. Yet, it is strikingly vulnerable to adversarial examples, e.g., word substitution attacks using only synonyms can easily fool a BERT-based sentiment analysis model. In this paper, we demonstrate that adversarial training, the prevalent defense technique, does not directly fit a conventional fine-tuning scenario, because it suffers severely from catastrophic forgetting: failing to retain the generic and robust linguistic features that have already been captured by the pre-trained model. In this light, we propose Robust Informative Fine-Tuning (RIFT), a novel adversarial fine-tuning method from an information-theoretical perspective. In particular, RIFT encourages an objective model to retain the features learned from the pre-trained model throughout the entire fine-tuning process, whereas a conventional one only uses the pre-trained weights for initialization. Experimental results show that RIFT consistently outperforms the state-of-the-arts on two popular NLP tasks: sentiment analysis and natural language inference, under different attacks across various pre-trained language models.
****

Recursive Bayesian Networks: Generalising and Unifying Probabilistic Context-Free Grammars and Dynamic Bayesian Networks

Robert Lieck, Martin Rohrmeier

Probabilistic context-free grammars (PCFGs) and dynamic Bayesian networks (DBNs) are widely used sequence models with complementary strengths and limitations. While PCFGs allow for nested hierarchical dependencies (tree structures), their latent variables (non-terminal symbols) have to be discrete. In contrast, DBNs allow for continuous latent variables, but the dependencies are strictly sequential (chain structure). Therefore, neither can be applied if the latent variables are assumed to be continuous and also to have a nested hierarchical dependency structure. In this paper, we present Recursive Bayesian Networks (RBNs), which generalise and unify PCFGs and DBNs, combining their strengths and containing both as special cases. RBNs define a joint distribution over tree-structured Bayesian networks with discrete or continuous latent variables. The main challenge lies in performing joint inference over the exponential number of possible structures and the continuous variables. We provide two solutions: 1) For arbitrary RBNs, we generalise inside and outside probabilities from PCFGs to the mixed discrete-continuous case, which allows for maximum posterior estimates of the continuous latent variables via gradient descent, while marginalising over network structures. 2) For Gaussian RBNs, we additionally derive an analytic approximation of the marginal data likelihood (evidence) and marginal posterior distribution, allow

ing for robust parameter optimisation and Bayesian inference. The capacity and diverse applications of RBNs are illustrated on two examples: In a quantitative evaluation on synthetic data, we demonstrate and discuss the advantage of RBNs for segmentation and tree induction from noisy sequences, compared to change point detection and hierarchical clustering. In an application to musical data, we approach the unsolved problem of hierarchical music analysis from the raw note level and compare our results to expert annotations.
****
EF21: A New, Simpler, Theoretically Better, and Practically Faster Error Feedback
Peter Richtarik, Igor Sokolov, Ilyas Fatkhullin
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Mixture weights optimisation for Alpha-Divergence Variational Inference
Kamélia Daudel, randal douc
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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-dependent Label-noise Learning under a Structural Causal Model
Yu Yao, Tongliang Liu, Mingming Gong, Bo Han, Gang Niu, 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-authors prior to requesting a name change in the electronic proceedings.
****
Combining Human Predictions with Model Probabilities via Confusion Matrices and Calibration
Gavin Kerrigan, Padhraic Smyth, Mark Steyvers
An increasingly common use case for machine learning models is augmenting the abilities of human decision makers. For classification tasks where neither the human nor model are perfectly accurate, a key step in obtaining high performance is combining their individual predictions in a manner that leverages their relative strengths. In this work, we develop a set of algorithms that combine the probabilistic output of a model with the class-level output of a human. We show theoretically that the accuracy of our combination model is driven not only by the individual human and model accuracies, but also by the model's confidence. Empirical results on image classification with CIFAR-10 and a subset of ImageNet demonstrate that such human-model combinations consistently have higher accuracies than the model or human alone, and that the parameters of the combination method can be estimated effectively with as few as ten labeled datapoints.
****
$\texttt{LeadCache}$: Regret-Optimal Caching in Networks
Debjit Paria, Abhishek Sinha
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Probabilistic Attention for Interactive Segmentation
Prasad Gabbur, Manjot Bilkhu, Javier Movellan
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

Influence Patterns for Explaining Information Flow in BERT
Kaiji Lu, Zifan Wang, Piotr Mardziel, Anupam Datta
While attention is all you need may be proving true, we do not know why: attention-based transformer models such as BERT are superior but how information flows from input tokens to output predictions are unclear. We introduce influence patterns, abstractions of sets of paths through a transformer model. Patterns quantify and localize the flow of information to paths passing through a sequence of model nodes. Experimentally, we find that significant portion of information flow in BERT goes through skip connections instead of attention heads. We further show that consistency of patterns across instances is an indicator of BERT’s performance. Finally, we demonstrate that patterns account for far more model performance than previous attention-based and layer-based methods.
****
Robust Regression Revisited: Acceleration and Improved Estimation Rates
Arun Jambulapati, Jerry Li, Tselil Schramm, Kevin Tian
We study fast algorithms for statistical regression problems under the strong contamination model, where the goal is to approximately optimize a generalized linear model (GLM) given adversarially corrupted samples. Prior works in this line of research were based on the \emph{robust gradient descent} framework of \cite{PrasadSBR20}, a first-order method using biased gradient queries, or the \emph{Sever} framework of \cite{DiakonikolasKK019}, an iterative outlier-removal method calling a stationary point finder. We present nearly-linear time algorithms for robust regression problems with improved runtime or estimation guarantees compared to the state-of-the-art. For the general case of smooth GLMs (e.g.\ logistic regression), we show that the robust gradient descent framework of \cite{PrasadSBR20} can be \emph{accelerated}, and show our algorithm extends to optimizing the Moreau envelopes of Lipschitz GLMs (e.g.\ support vector machines), answering several open questions in the literature. For the well-studied case of robust linear regression, we present an alternative approach obtaining improved estimation rates over prior nearly-linear time algorithms. Interestingly, our algorithm starts with an identifiability proof introduced in the context of the sum-of-squares algorithm of \cite{BakshiP21}, which achieved optimal error rates while requiring large polynomial runtime and sample complexity. We reinterpret their proof within the Sever framework and obtain a dramatically faster and more sample-efficient algorithm under fewer distributional assumptions.
****
Automatic Unsupervised Outlier Model Selection
Yue Zhao, Ryan Rossi, Leman Akoglu
Given an unsupervised outlier detection task on a new dataset, how can we automatically select a good outlier detection algorithm and its hyperparameter(s) (collectively called a model)? In this work, we tackle the unsupervised outlier model selection (UOMS) problem, and propose MetaOD, a principled, data-driven approach to UOMS based on meta-learning. The UOMS problem is notoriously challenging, as compared to model selection for classification and clustering, since (i) model evaluation is infeasible due to the lack of hold-out data with labels, and (ii) model comparison is infeasible due to the lack of a universal objective function. MetaOD capitalizes on the performances of a large body of detection models on historical outlier detection benchmark datasets, and carries over this prior experience to automatically select an effective model to be employed on a new dataset without any labels, model evaluations or model comparisons. To capture task similarity within our meta-learning framework, we introduce specialized meta-features that quantify outlying characteristics of a dataset. Extensive experiments show that selecting a model by MetaOD significantly outperforms no model selection (e.g. always using the same popular model or the ensemble of many) as well as other meta-learning techniques that we tailored for UOMS. Moreover upon (meta-)training, MetaOD is extremely efficient at test time; selecting from a large pool of 300+ models takes less than 1 second for a new task. We open-source MetaOD and our meta-learning database for practical use and to foster further research on the UOMS problem.
****

Pruning Randomly Initialized Neural Networks with Iterative Randomization
Daiki Chijiwa, Shin'ya Yamaguchi, Yasutoshi Ida, Kenji Umakoshi, Tomohiro INOUE
Pruning the weights of randomly initialized neural networks plays an important role in the context of lottery ticket hypothesis. Ramanujan et al. (2020) empirically showed that only pruning the weights can achieve remarkable performance instead of optimizing the weight values. However, to achieve the same level of performance as the weight optimization, the pruning approach requires more parameters in the networks before pruning and thus more memory space. To overcome this parameter inefficiency, we introduce a novel framework to prune randomly initialized neural networks with iteratively randomizing weight values (IteRand). Theoretically, we prove an approximation theorem in our framework, which indicates that the randomizing operations are provably effective to reduce the required number of the parameters. We also empirically demonstrate the parameter efficiency in multiple experiments on CIFAR-10 and ImageNet.
****
Probing Inter-modality: Visual Parsing with Self-Attention for Vision-and-Language Pre-training
Hongwei Xue, Yupan Huang, Bei Liu, Houwen Peng, Jianlong Fu, Houqiang Li, Jiebo Luo
Vision-Language Pre-training (VLP) aims to learn multi-modal representations from image-text pairs and serves for downstream vision-language tasks in a fine-tuning fashion. The dominant VLP models adopt a CNN-Transformer architecture, which embeds images with a CNN, and then aligns images and text with a Transformer. Visual relationship between visual contents plays an important role in image understanding and is the basic for inter-modal alignment learning. However, CNNs have limitations in visual relation learning due to local receptive field's weakness in modeling long-range dependencies. Thus the two objectives of learning visual relation and inter-modal alignment are encapsulated in the same Transformer network. Such design might restrict the inter-modal alignment learning in the Transformer by ignoring the specialized characteristic of each objective. To tackle this, we propose a fully Transformer visual embedding for VLP to better learn visual relation and further promote inter-modal alignment. Specifically, we propose a metric named Inter-Modality Flow (IMF) to measure the interaction between vision and language modalities (i.e., inter-modality). We also design a novel masking optimization mechanism named Masked Feature Regression (MFR) in Transformer to further promote the inter-modality learning. To the best of our knowledge, this is the first study to explore the benefit of Transformer for visual feature learning in VLP. We verify our method on a wide range of vision-language tasks, including Visual Question Answering (VQA), Visual Entailment and Visual Reasoning. Our approach not only outperforms the state-of-the-art VLP performance, but also shows benefits on the IMF metric.
****
Stability and Generalization of Bilevel Programming in Hyperparameter Optimization
Fan Bao, Guoqiang Wu, Chongxuan LI, Jun Zhu, Bo Zhang
The (gradient-based) bilevel programming framework is widely used in hyperparameter optimization and has achieved excellent performance empirically. Previous theoretical work mainly focuses on its optimization properties, while leaving the analysis on generalization largely open. This paper attempts to address the issue by presenting an expectation bound w.r.t. the validation set based on uniform stability. Our results can explain some mysterious behaviours of the bilevel programming in practice, for instance, overfitting to the validation set. We also present an expectation bound for the classical cross-validation algorithm. Our results suggest that gradient-based algorithms can be better than cross-validation under certain conditions in a theoretical perspective. Furthermore, we prove that regularization terms in both the outer and inner levels can relieve the overfitting problem in gradient-based algorithms. In experiments on feature learning and data reweighting for noisy labels, we corroborate our theoretical findings.
****
Regime Switching Bandits

Xiang Zhou, Yi Xiong, Ningyuan Chen, Xuefeng GAO
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
MixACM: Mixup-Based Robustness Transfer via Distillation of Activated Channel Maps
Awais Muhammad, Fengwei Zhou, Chuanlong Xie, Jiawei Li, Sung-Ho Bae, Zhenguo Li
Deep neural networks are susceptible to adversarially crafted, small, and imperceptible changes in the natural inputs. The most effective defense mechanism against these examples is adversarial training which constructs adversarial examples during training by iterative maximization of loss. The model is then trained to minimize the loss on these constructed examples. This min-max optimization requires more data, larger capacity models, and additional computing resources. It also degrades the standard generalization performance of a model. Can we achieve robustness more efficiently? In this work, we explore this question from the perspective of knowledge transfer. First, we theoretically show the transferability of robustness from an adversarially trained teacher model to a student model with the help of mixup augmentation. Second, we propose a novel robustness transfer method called Mixup-Based Activated Channel Maps (MixACM) Transfer. MixACM transfers robustness from a robust teacher to a student by matching activated channel maps generated without expensive adversarial perturbations. Finally, extensive experiments on multiple datasets and different learning scenarios show our method can transfer robustness while also improving generalization on natural images.
****
Localization, Convexity, and Star Aggregation
Suhas Vijaykumar
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Aligning Silhouette Topology for Self-Adaptive 3D Human Pose Recovery
Ramesha Rakesh Mugaludi, Jogendra Nath Kundu, Varun Jampani, Venkatesh Babu R
Articulation-centric 2D/3D pose supervision forms the core training objective in most existing 3D human pose estimation techniques. Except for synthetic source environments, acquiring such rich supervision for each real target domain at deployment is highly inconvenient. However, we realize that standard foreground silhouette estimation techniques (on static camera feeds) remain unaffected by domain-shifts. Motivated by this, we propose a novel target adaptation framework that relies only on silhouette supervision to adapt a source-trained model-based regressor. However, in the absence of any auxiliary cue (multi-view, depth, or 2D pose), an isolated silhouette loss fails to provide a reliable pose-specific gradient and requires to be employed in tandem with a topology-centric loss. To this end, we develop a series of convolution-friendly spatial transformations in order to disentangle a topological-skeleton representation from the raw silhouette. Such a design paves the way to devise a Chamfer-inspired spatial topological-alignment loss via distance field computation, while effectively avoiding any gradient hindering spatial-to-pointset mapping. Experimental results demonstrate our superiority against prior-arts in self-adapting a source trained model to diverse unlabeled target domains, such as a) in-the-wild datasets, b) low-resolution image domains, and c) adversarially perturbed image domains (via UAP).
****
Self-Adaptable Point Processes with Nonparametric Time Decays
Zhimeng Pan, Zheng Wang, Jeff M Phillips, Shandian Zhe
Requests for name changes in the electronic proceedings will be accepted with 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.
****
Offline Meta Reinforcement Learning -- Identifiability Challenges and Effective Data Collection Strategies
Ron Dorfman, Idan Shenfeld, Aviv Tamar
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
RoMA: Robust Model Adaptation for Offline Model-based Optimization
Sihyun Yu, Sungsoo Ahn, Le Song, Jinwoo Shin
We consider the problem of searching an input maximizing a black-box objective function given a static dataset of input-output queries. A popular approach to solving this problem is maintaining a proxy model, e.g., a deep neural network (DNN), that approximates the true objective function. Here, the main challenge is how to avoid adversarially optimized inputs during the search, i.e., the inputs where the DNN highly overestimates the true objective function. To handle the issue, we propose a new framework, coined robust model adaptation (RoMA), based on gradient-based optimization of inputs over the DNN. Specifically, it consists of two steps: (a) a pre-training strategy to robustly train the proxy model and (b) a novel adaptation procedure of the proxy model to have robust estimates for a specific set of candidate solutions. At a high level, our scheme utilizes the local smoothness prior to overcome the brittleness of the DNN. Experiments under various tasks show the effectiveness of RoMA compared with previous methods, obtaining state-of-the-art results, e.g., RoMA outperforms all at 4 out of 6 tasks and achieves runner-up results at the remaining tasks.
****
Flexible Option Learning
Martin Klissarov, Doina Precup
Temporal abstraction in reinforcement learning (RL), offers the promise of improving generalization and knowledge transfer in complex environments, by propagating information more efficiently over time. Although option learning was initially formulated in a way that allows updating many options simultaneously, using off-policy, intra-option learning (Sutton, Precup & Singh, 1999) , many of the recent hierarchical reinforcement learning approaches only update a single option at a time: the option currently executing. We revisit and extend intra-option learning in the context of deep reinforcement learning, in order to enable updating all options consistent with current primitive action choices, without introducing any additional estimates. Our method can therefore be naturally adopted in most hierarchical RL frameworks. When we combine our approach with the option-critic algorithm for option discovery, we obtain significant improvements in performance and data-efficiency across a wide variety of domains.
****
Faster Directional Convergence of Linear Neural Networks under Spherically Symmetric Data
Dachao Lin, Ruoyu Sun, Zhihua Zhang
In this paper, we study gradient methods for training deep linear neural networks with binary cross-entropy loss. In particular, we show global directional convergence guarantees from a polynomial rate to a linear rate for (deep) linear networks with spherically symmetric data distribution, which can be viewed as a specific zero-margin dataset. Our results do not require the assumptions in other works such as small initial loss, presumed convergence of weight direction, or overparameterization. We also characterize our findings in experiments.
****
Online Facility Location with Multiple Advice
Matteo Almanza, Flavio Chierichetti, Silvio Lattanzi, Alessandro Panconesi, Giuseppe Re
Clustering is a central topic in unsupervised learning and its online formulation has received a lot of attention in recent years. In this paper, we study the c

lassic facility location problem in the presence of multiple machine-learned adv
ice. We design an algorithm with provable performance guarantees such that, if t
he advice is good, it outperforms the best-known online algorithms for the probl
em, and if it is bad it still matches their performance.We complement our theore
tical analysis with an in-depth study of the performance of our algorithm, showi
ng its effectiveness on synthetic and real-world data sets.
****

Credit Assignment in Neural Networks through Deep Feedback Control
Alexander Meulemans, Matilde Tristany Farinha, Javier Garcia Ordonez, Pau Vilime
lis Aceituno, João Sacramento, Benjamin F. Grewe
The success of deep learning sparked interest in whether the brain learns by usi
ng similar techniques for assigning credit to each synaptic weight for its contr
ibution to the network output. However, the majority of current attempts at biol
ogically-plausible learning methods are either non-local in time, require highly
 specific connectivity motifs, or have no clear link to any known mathematical o
ptimization method. Here, we introduce Deep Feedback Control (DFC), a new learni
ng method that uses a feedback controller to drive a deep neural network to matc
h a desired output target and whose control signal can be used for credit assign
ment. The resulting learning rule is fully local in space and time and approxima
tes Gauss-Newton optimization for a wide range of feedback connectivity patterns
. To further underline its biological plausibility, we relate DFC to a multi-com
partment model of cortical pyramidal neurons with a local voltage-dependent syna
ptic plasticity rule, consistent with recent theories of dendritic processing. B
y combining dynamical system theory with mathematical optimization theory, we pr
ovide a strong theoretical foundation for DFC that we corroborate with detailed
results on toy experiments and standard computer-vision benchmarks.
****

Robust Online Correlation Clustering
Silvio Lattanzi, Benjamin Moseley, Sergei Vassilvitskii, Yuyan Wang, Rudy Zhou
In correlation clustering we are given a set of points along with recommendation
s whether each pair of points should be placed in the same cluster or into separ
ate clusters. The goal cluster the points to minimize disagreements from the rec
ommendations. We study the correlation clustering problem in the online setting.
, where points arrive one at a time, and upon arrival the algorithm must make an
 irrevocable cluster assignment decision. While the online version is natural, t
here is a simple lower bound that rules out any algorithm with a non-trivial com
petitive ratio. In this work we go beyond worst case analysis, and show that the
 celebrated Pivot algorithm performs well when given access to a small number of
 random samples from the input. Moreover, we prove that Pivot is robust to addit
ional adversarial perturbations of the sample set in this setting. We conclude w
ith an empirical analysis validating our theoretical findings.
****

Neural Additive Models: Interpretable Machine Learning with Neural Nets
Rishabh Agarwal, Levi Melnick, Nicholas Frosst, Xuezhou Zhang, Ben Lengerich, Ri
ch Caruana, Geoffrey E. Hinton
Deep neural networks (DNNs) are powerful black-box predictors that have achieved
 impressive performance on a wide variety of tasks. However, their accuracy come
s at the cost of intelligibility: it is usually unclear how they make their deci
sions. This hinders their applicability to high stakes decision-making domains s
uch as healthcare. We propose Neural Additive Models (NAMs) which combine some o
f the expressivity of DNNs with the inherent intelligibility of generalized addi
tive models. NAMs learn a linear combination of neural networks that each attend
 to a single input feature. These networks are trained jointly and can learn arb
itrarily complex relationships between their input feature and the output. Our e
xperiments on regression and classification datasets show that NAMs are more acc
urate than widely used intelligible models such as logistic regression and shall
ow decision trees. They perform similarly to existing state-of-the-art generaliz
ed additive models in accuracy, but are more flexible because they are based on
neural nets instead of boosted trees. To demonstrate this, we show how NAMs can
be used for multitask learning on synthetic data and on the COMPAS recidivism da

ta due to their composability, and demonstrate that the differentiability of NAM
s allows them to train more complex interpretable models for COVID-19.
****
Representation Learning for Event-based Visuomotor Policies
Sai Vemprala, Sami Mian, Ashish Kapoor
Event-based cameras are dynamic vision sensors that provide asynchronous measure
ments of changes in per-pixel brightness at a microsecond level. This makes them
 significantly faster than conventional frame-based cameras, and an appealing ch
oice for high-speed robot navigation. While an interesting sensor modality, this
 asynchronously streamed event data poses a challenge for machine learning based
 computer vision techniques that are more suited for synchronous, frame-based da
ta. In this paper, we present an event variational autoencoder through which com
pact representations can be learnt directly from asynchronous spatiotemporal eve
nt data. Furthermore, we show that such pretrained representations can be used f
or event-based reinforcement learning instead of end-to-end reward driven percep
tion. We validate this framework of learning event-based visuomotor policies by
applying it to an obstacle avoidance scenario in simulation. Compared to techniq
ues that treat event data as images, we show that representations learnt from ev
ent streams result in faster policy training, adapt to different control capacit
ies, and demonstrate a higher degree of robustness to environmental changes and
sensor noise.
****
Kernel Functional Optimisation
Arun Kumar Anjanapura Venkatesh, Alistair Shilton, Santu Rana, Sunil Gupta, Svet
ha Venkatesh
Traditional methods for kernel selection rely on parametric kernel functions or
a combination thereof and although the kernel hyperparameters are tuned, these m
ethods often provide sub-optimal results due to the limitations induced by the p
arametric forms. In this paper, we propose a novel formulation for kernel select
ion using efficient Bayesian optimisation to find the best fitting non-parametri
c kernel. The kernel is expressed using a linear combination of functions sample
d from a prior Gaussian Process (GP) defined by a hyperkernel. We also provide a
 mechanism to ensure the positive definiteness of the Gram matrix constructed us
ing the resultant kernels. Our experimental results on GP regression and Support
 Vector Machine (SVM) classification tasks involving both synthetic functions an
d several real-world datasets show the superiority of our approach over the stat
e-of-the-art.
****
Generalized Shape Metrics on Neural Representations
Alex H Williams, Erin Kunz, Simon Kornblith, Scott Linderman
Understanding the operation of biological and artificial networks remains a diff
icult and important challenge. To identify general principles, researchers are i
ncreasingly interested in surveying large collections of networks that are train
ed on, or biologically adapted to, similar tasks. A standardized set of analysis
 tools is now needed to identify how network-level covariates---such as architec
ture, anatomical brain region, and model organism---impact neural representation
s (hidden layer activations). Here, we provide a rigorous foundation for these a
nalyses by defining a broad family of metric spaces that quantify representation
al dissimilarity. Using this framework, we modify existing representational simi
larity measures based on canonical correlation analysis and centered kernel alig
nment to satisfy the triangle inequality, formulate a novel metric that respects
 the inductive biases in convolutional layers, and identify approximate Euclidea
n embeddings that enable network representations to be incorporated into essenti
ally any off-the-shelf machine learning method. We demonstrate these methods on
large-scale datasets from biology (Allen Institute Brain Observatory) and deep l
earning (NAS-Bench-101). In doing so, we identify relationships between neural r
epresentations that are interpretable in terms of anatomical features and model
performance.
****
Diverse Message Passing for Attribute with Heterophily

Liang Yang, Mengzhe Li, Liyang Liu, bingxin niu, Chuan Wang, Xiaochun Cao, Yuanfang Guo
Most of the existing GNNs can be modeled via the Uniform Message Passing framework. This framework considers all the attributes of each node in its entirety, shares the uniform propagation weights along each edge, and focuses on the uniform weight learning. The design of this framework possesses two prerequisites, the simplification of homophily and heterophily to the node-level property and the ignorance of attribute differences. Unfortunately, different attributes possess diverse characteristics. In this paper, the network homophily rate defined with respect to the node labels is extended to attribute homophily rate by taking the attributes as weak labels. Based on this attribute homophily rate, we propose a Diverse Message Passing (DMP) framework, which specifies every attribute propagation weight on each edge. Besides, we propose two specific strategies to significantly reduce the computational complexity of DMP to prevent the overfitting issue. By investigating the spectral characteristics, existing spectral GNNs are actually equivalent to a degenerated version of DMP. From the perspective of numerical optimization, we provide a theoretical analysis to demonstrate DMP's powerful representation ability and the ability of alleviating the over-smoothing issue. Evaluations on various real networks demonstrate the superiority of our DMP on handling the networks with heterophily and alleviating the over-smoothing issue, compared to the existing state-of-the-arts.
****
Towards Robust Bisimulation Metric Learning
Mete Kemertas, Tristan Aumentado-Armstrong
Learned representations in deep reinforcement learning (DRL) have to extract task-relevant information from complex observations, balancing between robustness to distraction and informativeness to the policy. Such stable and rich representations, often learned via modern function approximation techniques, can enable practical application of the policy improvement theorem, even in high-dimensional continuous state-action spaces. Bisimulation metrics offer one solution to this representation learning problem, by collapsing functionally similar states together in representation space, which promotes invariance to noise and distractors. In this work, we generalize value function approximation bounds for on-policy bisimulation metrics to non-optimal policies and approximate environment dynamics. Our theoretical results help us identify embedding pathologies that may occur in practical use. In particular, we find that these issues stem from an underconstrained dynamics model and an unstable dependence of the embedding norm on the reward signal in environments with sparse rewards. Further, we propose a set of practical remedies: (i) a norm constraint on the representation space, and (ii) an extension of prior approaches with intrinsic rewards and latent space regularization. Finally, we provide evidence that the resulting method is not only more robust to sparse reward functions, but also able to solve challenging continuous control tasks with observational distractions, where prior methods fail.
****
Beyond BatchNorm: Towards a Unified Understanding of Normalization in Deep Learning
Ekdeep S Lubana, Robert Dick, Hidenori Tanaka
Inspired by BatchNorm, there has been an explosion of normalization layers in deep learning. Recent works have identified a multitude of beneficial properties in BatchNorm to explain its success. However, given the pursuit of alternative normalization layers, these properties need to be generalized so that any given layer's success/failure can be accurately predicted. In this work, we take a first step towards this goal by extending known properties of BatchNorm in randomly initialized deep neural networks (DNNs) to several recently proposed normalization layers. Our primary findings follow: (i) similar to BatchNorm, activations-based normalization layers can prevent exponential growth of activations in ResNets, but parametric techniques require explicit remedies; (ii) use of GroupNorm can ensure an informative forward propagation, with different samples being assigned dissimilar activations, but increasing group size results in increasingly indistinguishable activations for different samples, explaining slow convergence spe

ed in models with LayerNorm; and (iii) small group sizes result in large gradien t norm in earlier layers, hence explaining training instability issues in Instan ce Normalization and illustrating a speed-stability tradeoff in GroupNorm. Overa ll, our analysis reveals a unified set of mechanisms that underpin the success o f normalization methods in deep learning, providing us with a compass to systema tically explore the vast design space of DNN normalization layers.
****
Representation Learning Beyond Linear Prediction Functions
Ziping Xu, Ambuj Tewari
Recent papers on the theory of representation learning has shown the importance of a quantity called diversity when generalizing from a set of source tasks to a target task. Most of these papers assume that the function mapping shared repre sentations to predictions is linear, for both source and target tasks. In practi ce, researchers in deep learning use different numbers of extra layers following the pretrained model based on the difficulty of the new task. This motivates us to ask whether diversity can be achieved when source tasks and the target task use different prediction function spaces beyond linear functions. We show that d iversity holds even if the target task uses a neural network with multiple layer s, as long as source tasks use linear functions. If source tasks use nonlinear p rediction functions, we provide a negative result by showing that depth-1 neural networks with ReLu activation function need exponentially many source tasks to achieve diversity. For a general function class, we find that eluder dimension g ives a lower bound on the number of tasks required for diversity. Our theoretica l results imply that simpler tasks generalize better. Though our theoretical res ults are shown for the global minimizer of empirical risks, their qualitative pr edictions still hold true for gradient-based optimization algorithms as verified by our simulations on deep neural networks.
****
Volume Rendering of Neural Implicit Surfaces
Lior Yariv, Jiatao Gu, Yoni Kasten, Yaron Lipman
Neural volume rendering became increasingly popular recently due to its success in synthesizing novel views of a scene from a sparse set of input images. So far , the geometry learned by neural volume rendering techniques was modeled using a generic density function. Furthermore, the geometry itself was extracted using an arbitrary level set of the density function leading to a noisy, often low fid elity reconstruction.The goal of this paper is to improve geometry representatio n and reconstruction in neural volume rendering. We achieve that by modeling the volume density as a function of the geometry. This is in contrast to previous w ork modeling the geometry as a function of the volume density. In more detail, w e define the volume density function as Laplace's cumulative distribution functi on (CDF) applied to a signed distance function (SDF) representation. This simple density representation has three benefits: (i) it provides a useful inductive b ias to the geometry learned in the neural volume rendering process; (ii) it faci litates a bound on the opacity approximation error, leading to an accurate sampl ing of the viewing ray. Accurate sampling is important to provide a precise coup ling of geometry and radiance; and (iii) it allows efficient unsupervised disent anglement of shape and appearance in volume rendering.Applying this new density representation to challenging scene multiview datasets produced high quality geo metry reconstructions, outperforming relevant baselines. Furthermore, switching shape and appearance between scenes is possible due to the disentanglement of th e two.
****
MAUVE: Measuring the Gap Between Neural Text and Human Text using Divergence Fro ntiers
Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Wellec k, Yejin Choi, Zaid Harchaoui
As major progress is made in open-ended text generation, measuring how close mac hine-generated text is to human language remains a critical open problem. We int roduce Mauve, a comparison measure for open-ended text generation, which directl y compares the learnt distribution from a text generation model to the distribut

ion of human-written text using divergence frontiers. Mauve scales up to modern text generation models by computing information divergences in a quantized embedding space. Through an extensive empirical study on three open-ended generation tasks, we find that Mauve identifies known properties of generated text, scales naturally with model size, and correlates with human judgments, with fewer restrictions than existing distributional evaluation metrics.
****
Accurately Solving Rod Dynamics with Graph Learning
Han Shao, Tassilo Kugelstadt, Torsten Hädrich, Wojtek Palubicki, Jan Bender, Soeren Pirk, Dominik L Michels
Iterative solvers are widely used to accurately simulate physical systems. These solvers require initial guesses to generate a sequence of improving approximate solutions. In this contribution, we introduce a novel method to accelerate iterative solvers for rod dynamics with graph networks (GNs) by predicting the initial guesses to reduce the number of iterations. Unlike existing methods that aim to learn physical systems in an end-to-end manner, our approach guarantees long-term stability and therefore leads to more accurate solutions. Furthermore, our method improves the run time performance of traditional iterative solvers for rod dynamics. To explore our method we make use of position-based dynamics (PBD) as a common solver for physical systems and evaluate it by simulating the dynamics of elastic rods. Our approach is able to generalize across different initial conditions, discretizations, and realistic material properties. We demonstrate that it also performs well when taking discontinuous effects into account such as collisions between individual rods. Finally, to illustrate the scalability of our approach, we simulate complex 3D tree models composed of over a thousand individual branch segments swaying in wind fields.
****
Limiting fluctuation and trajectorial stability of multilayer neural networks with mean field training
Huy Tuan Pham, Phan-Minh Nguyen
The mean field theory of multilayer neural networks centers around a particular infinite-width scaling, in which the learning dynamics is shown to be closely tracked by the mean field limit. A random fluctuation around this infinite-width limit is expected from a large-width expansion to the next order. This fluctuation has been studied only in the case of shallow networks, where previous works employ heavily technical notions or additional formulation ideas amenable only to that case. Treatment of the multilayer case has been missing, with the chief difficulty in finding a formulation that must capture the stochastic dependency across not only time but also depth.In this work, we initiate the study of the fluctuation in the case of multilayer networks, at any network depth. Leveraging on the neuronal embedding framework recently introduced by Nguyen and Pham, we systematically derive a system of dynamical equations, called the second-order mean field limit, that captures the limiting fluctuation distribution. We demonstrate through the framework the complex interaction among neurons in this second-order mean field limit, the stochasticity with cross-layer dependency and the nonlinear time evolution inherent in the limiting fluctuation. A limit theorem is proven to relate quantitatively this limit to the fluctuation realized by large-width networks.We apply the result to show a stability property of gradient descent mean field training: in the large-width regime, along the training trajectory, it progressively biases towards a solution with "minimal fluctuation" (in fact, vanishing fluctuation) in the learned output function, even after the network has been initialized at or has converged (sufficiently fast) to a global optimum. This extends a similar phenomenon previously shown only for shallow networks with a squared loss in the empirical risk minimization setting, to multilayer networks with a loss function that is not necessarily convex in a more general setting.
****
Medical Dead-ends and Learning to Identify High-Risk States and Treatments
Mehdi Fatemi, Taylor W. Killian, Jayakumar Subramanian, Marzyeh Ghassemi
Machine learning has successfully framed many sequential decision making problem

s as either supervised prediction, or optimal decision-making policy identification via reinforcement learning. In data-constrained offline settings, both approaches may fail as they assume fully optimal behavior or rely on exploring alternatives that may not exist. We introduce an inherently different approach that identifies "dead-ends" of a state space. We focus on patient condition in the intensive care unit, where a "medical dead-end" indicates that a patient will expire, regardless of all potential future treatment sequences. We postulate "treatment security" as avoiding treatments with probability proportional to their chance of leading to dead-ends, present a formal proof, and frame discovery as an RL problem. We then train three independent deep neural models for automated state construction, dead-end discovery and confirmation. Our empirical results discover that dead-ends exist in real clinical data among septic patients, and further reveal gaps between secure treatments and those administered.
****
Overcoming the Convex Barrier for Simplex Inputs
Harkirat Singh Behl, M. Pawan Kumar, Philip Torr, Krishnamurthy Dvijotham
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
High-probability Bounds for Non-Convex Stochastic Optimization with Heavy Tails
Ashok Cutkosky, Harsh Mehta
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Batch Normalization Orthogonalizes Representations in Deep Random Networks
Hadi Daneshmand, Amir Joudaki, Francis Bach
This paper underlines an elegant property of batch-normalization (BN): Successive batch normalizations with random linear updates make samples increasingly orthogonal. We establish a non-asymptotic characterization of the interplay between depth, width, and the orthogonality of deep representations. More precisely, we prove, under a mild assumption, the deviation of the representations from orthogonality rapidly decays with depth up to a term inversely proportional to the network width. This result has two main theoretical and practical implications: 1) Theoretically, as the depth grows, the distribution of the outputs contracts to a Wasserstein-2 ball around an isotropic normal distribution. Furthermore, the radius of this Wasserstein ball shrinks with the width of the network. 2) Practically, the orthogonality of the representations directly influences the performance of stochastic gradient descent (SGD). When representations are initially aligned, we observe SGD wastes many iterations to disentangle representations before the classification. Nevertheless, we experimentally show that starting optimization from orthogonal representations is sufficient to accelerate SGD, with no need for BN.
****
Support vector machines and linear regression coincide with very high-dimensional features
Navid Ardeshir, Clayton Sanford, Daniel J. Hsu
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Coupled Segmentation and Edge Learning via Dynamic Graph Propagation
Zhiding Yu, Rui Huang, Wonmin Byeon, Sifei Liu, Guilin Liu, Thomas Breuel, Anima Anandkumar, Jan Kautz
Image segmentation and edge detection are both central problems in perceptual grouping. It is therefore interesting to study how these two tasks can be coupled

to benefit each other. Indeed, segmentation can be easily transformed into contour edges to guide edge learning. However, the converse is nontrivial since general edges may not always form closed contours. In this paper, we propose a principled end-to-end framework for coupled edge and segmentation learning, where edges are leveraged as pairwise similarity cues to guide segmentation. At the core of our framework is a recurrent module termed as dynamic graph propagation (DGP) layer that performs message passing on dynamically constructed graphs. The layer uses learned gating to dynamically select neighbors for message passing using max-pooling. The output from message passing is further gated with an edge signal to refine segmentation. Experiments demonstrate that the proposed framework is able to let both tasks mutually improve each other. On Cityscapes validation, our best model achieves 83.7% mIoU in semantic segmentation and 78.7% maximum F-score in semantic edge detection. Our method also leads to improved zero-shot robustness on Cityscapes with natural corruptions (Cityscapes-C).
****
Offline RL Without Off-Policy Evaluation
David Brandfonbrener, Will Whitney, Rajesh Ranganath, Joan Bruna
Most prior approaches to offline reinforcement learning (RL) have taken an iterative actor-critic approach involving off-policy evaluation. In this paper we show that simply doing one step of constrained/regularized policy improvement using an on-policy Q estimate of the behavior policy performs surprisingly well. This one-step algorithm beats the previously reported results of iterative algorithms on a large portion of the D4RL benchmark. The one-step baseline achieves this strong performance while being notably simpler and more robust to hyperparameters than previously proposed iterative algorithms. We argue that the relatively poor performance of iterative approaches is a result of the high variance inherent in doing off-policy evaluation and magnified by the repeated optimization of policies against those estimates. In addition, we hypothesize that the strong performance of the one-step algorithm is due to a combination of favorable structure in the environment and behavior policy.
****
Continuous vs. Discrete Optimization of Deep Neural Networks
Omer Elkabetz, Nadav Cohen
Existing analyses of optimization in deep learning are either continuous, focusing on (variants of) gradient flow, or discrete, directly treating (variants of) gradient descent. Gradient flow is amenable to theoretical analysis, but is stylized and disregards computational efficiency. The extent to which it represents gradient descent is an open question in the theory of deep learning. The current paper studies this question. Viewing gradient descent as an approximate numerical solution to the initial value problem of gradient flow, we find that the degree of approximation depends on the curvature around the gradient flow trajectory. We then show that over deep neural networks with homogeneous activations, gradient flow trajectories enjoy favorable curvature, suggesting they are well approximated by gradient descent. This finding allows us to translate an analysis of gradient flow over deep linear neural networks into a guarantee that gradient descent efficiently converges to global minimum almost surely under random initialization. Experiments suggest that over simple deep neural networks, gradient descent with conventional step size is indeed close to gradient flow. We hypothesize that the theory of gradient flows will unravel mysteries behind deep learning.
****
CrypTen: Secure Multi-Party Computation Meets Machine Learning
Brian Knott, Shobha Venkataraman, Awni Hannun, Shubho Sengupta, Mark Ibrahim, Laurens van der Maaten
Secure multi-party computation (MPC) allows parties to perform computations on data while keeping that data private. This capability has great potential for machine-learning applications: it facilitates training of machine-learning models on private data sets owned by different parties, evaluation of one party's private model using another party's private data, etc. Although a range of studies implement machine-learning models via secure MPC, such implementations are not yet

mainstream. Adoption of secure MPC is hampered by the absence of flexible softwa
re frameworks that `"speak the language" of machine-learning researchers and eng
ineers. To foster adoption of secure MPC in machine learning, we present CrypTen
: a software framework that exposes popular secure MPC primitives via abstractio
ns that are common in modern machine-learning frameworks, such as tensor computa
tions, automatic differentiation, and modular neural networks. This paper descri
bes the design of CrypTen and measure its performance on state-of-the-art models
 for text classification, speech recognition, and image classification. Our benc
hmarks show that CrypTen's GPU support and high-performance communication betwee
n (an arbitrary number of) parties allows it to perform efficient private evalua
tion of modern machine-learning models under a semi-honest threat model. For exa
mple, two parties using CrypTen can securely predict phonemes in speech recordin
gs using Wav2Letter faster than real-time. We hope that CrypTen will spur adopti
on of secure MPC in the machine-learning community.
****
Can contrastive learning avoid shortcut solutions?
Joshua Robinson, Li Sun, Ke Yu, Kayhan Batmanghelich, Stefanie Jegelka, Suvrit S
ra
The generalization of representations learned via contrastive learning depends c
rucially on what features of the data are extracted. However, we observe that th
e contrastive loss does not always sufficiently guide which features are extract
ed, a behavior that can negatively impact the performance on downstream tasks vi
a “shortcuts", i.e., by inadvertently suppressing important predictive features.
  We find that feature extraction is influenced by the difficulty of the so-cal
led instance discrimination task (i.e., the task of discriminating pairs of simi
lar points from pairs of dissimilar ones). Although harder pairs improve the rep
resentation of some features, the improvement comes at the cost of suppressing p
reviously well represented features. In response, we propose implicit feature mo
dification (IFM), a method for altering positive and negative samples in order t
o guide contrastive models towards capturing a wider variety of predictive featu
res. Empirically, we observe that IFM reduces feature suppression, and as a resu
lt improves performance on vision and medical imaging tasks.
****
See More for Scene: Pairwise Consistency Learning for Scene Classification
Gongwei Chen, Xinhang Song, Bohan Wang, Shuqiang Jiang
Scene classification is a valuable classification subtask and has its own charac
teristics which still needs more in-depth studies. Basically, scene characterist
ics are distributed over the whole image, which cause the need of “seeing” compr
ehensive and informative regions. Previous works mainly focus on region discover
y and aggregation, while rarely involves the inherent properties of CNN along wi
th its potential ability to satisfy the requirements of scene classification. In
 this paper, we propose to understand scene images and the scene classification
CNN models in terms of the focus area. From this new perspective, we find that l
arge focus area is preferred in scene classification CNN models as a consequence
 of learning scene characteristics. Meanwhile, the analysis about existing train
ing schemes helps us to understand the effects of focus area, and also raises th
e question about optimal training method for scene classification. Pursuing the
better usage of scene characteristics, we propose a new learning scheme with a t
ailored loss in the goal of activating larger focus area on scene images. Since
the supervision of the target regions to be enlarged is usually lacked, our alte
rnative learning scheme is to erase already activated area, and allow the CNN mo
dels to activate more area during training. The proposed scheme is implemented b
y keeping the pairwise consistency between the output of the erased image and i
ts original one. In particular, a tailored loss is proposed to keep such pairwis
e consistency by leveraging category-relevance information. Experiments on Place
s365 show the significant improvements of our method with various CNNs. Our meth
od shows an inferior result on the object-centric dataset, ImageNet, which exper
imentally indicates that it captures the unique characteristics of scenes.
****
Provable Guarantees for Self-Supervised Deep Learning with Spectral Contrastive

Loss
Jeff Z. HaoChen, Colin Wei, Adrien Gaidon, Tengyu Ma
Recent works in self-supervised learning have advanced the state-of-the-art by relying on the contrastive learning paradigm, which learns representations by pushing positive pairs, or similar examples from the same class, closer together while keeping negative pairs far apart. Despite the empirical successes, theoretical foundations are limited -- prior analyses assume conditional independence of the positive pairs given the same class label, but recent empirical applications use heavily correlated positive pairs (i.e., data augmentations of the same image). Our work analyzes contrastive learning without assuming conditional independence of positive pairs using a novel concept of the augmentation graph on data. Edges in this graph connect augmentations of the same data, and ground-truth classes naturally form connected sub-graphs. We propose a loss that performs spectral decomposition on the population augmentation graph and can be succinctly written as a contrastive learning objective on neural net representations. Minimizing this objective leads to features with provable accuracy guarantees under linear probe evaluation. By standard generalization bounds, these accuracy guarantees also hold when minimizing the training contrastive loss. In all, this work provides the first provable analysis for contrastive learning where the guarantees can apply to realistic empirical settings.
****
Greedy Approximation Algorithms for Active Sequential Hypothesis Testing
Kyra Gan, Su Jia, Andrew 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.
****
When False Positive is Intolerant: End-to-End Optimization with Low FPR for Multipartite Ranking
Peisong Wen, Qianqian Xu, Zhiyong Yang, Yuan He, Qingming Huang
Multipartite ranking is a basic task in machine learning, where the Area Under the receiver operating characteristics Curve (AUC) is generally applied as the evaluation metric. Despite that AUC reflects the overall performance of the model, it is inconsistent with the expected performance in some application scenarios, where only a low False Positive Rate (FPR) is meaningful. To leverage high performance under low FPRs, we consider an alternative metric for multipartite ranking evaluating the True Positive Rate (TPR) at a given FPR, denoted as TPR@FPR. Unfortunately, the key challenge of direct TPR@FPR optimization is two-fold: \textbf{a)} the original objective function is not differentiable, making gradient backpropagation impossible; \textbf{b)} the loss function could not be written as a sum of independent instance-wise terms, making mini-batch based optimization infeasible.   To address these issues, we propose a novel framework on top of the deep learning framework named \textit{Cross-Batch Approximation for Multipartite Ranking (CBA-MR)}. In face of \textbf{a)}, we propose a differentiable surrogate optimization problem where the instances having a short-time effect on FPR are rendered with different weights based on the random walk hypothesis. To tackle \textbf{b)}, we propose a fast ranking estimation method, where the full-batch loss evaluation is replaced by a delayed update scheme with the help of an embedding cache. Finally, experimental results on four real-world benchmarks are provided to demonstrate the effectiveness of the proposed method.
****
Convex Polytope Trees
Mohammadreza Armandpour, Ali Sadeghian, Mingyuan Zhou
A decision tree is commonly restricted to use a single hyperplane to split the covariate space at each of its internal nodes. It often requires a large number of nodes to achieve high accuracy. In this paper, we propose convex polytope trees (CPT) to expand the family of decision trees by an interpretable generalization of their decision boundary. The splitting function at each node of CPT is based on the logical disjunction of a community of differently weighted probabilisti

c linear decision-makers, which also geometrically corresponds to a convex polytope in the covariate space. We use a nonparametric Bayesian prior at each node to infer the community's size, encouraging simpler decision boundaries by shrinking the number of polytope facets. We develop a greedy method to efficiently construct CPT and scalable end-to-end training algorithms for the tree parameters when the tree structure is given. We empirically demonstrate the efficiency of CPT over existing state-of-the-art decision trees in several real-world classification and regression tasks from diverse domains.
****
The Skellam Mechanism for Differentially Private Federated Learning
Naman Agarwal, Peter Kairouz, Ziyu Liu
We introduce the multi-dimensional Skellam mechanism, a discrete differential privacy mechanism based on the difference of two independent Poisson random variables. To quantify its privacy guarantees, we analyze the privacy loss distribution via a numerical evaluation and provide a sharp bound on the Rényi divergence between two shifted Skellam distributions. While useful in both centralized and distributed privacy applications, we investigate how it can be applied in the context of federated learning with secure aggregation under communication constraints. Our theoretical findings and extensive experimental evaluations demonstrate that the Skellam mechanism provides the same privacy-accuracy trade-offs as the continuous Gaussian mechanism, even when the precision is low. More importantly, Skellam is closed under summation and sampling from it only requires sampling from a Poisson distribution -- an efficient routine that ships with all machine learning and data analysis software packages. These features, along with its discrete nature and competitive privacy-accuracy trade-offs, make it an attractive practical alternative to the newly introduced discrete Gaussian mechanism.
****
Stability and Deviation Optimal Risk Bounds with Convergence Rate $O(1/n)$
Yegor Klochkov, Nikita Zhivotovskiy
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
SketchGen: Generating Constrained CAD Sketches
Wamiq Para, Shariq Bhat, Paul Guerrero, Tom Kelly, Niloy Mitra, Leonidas J. Guibas, Peter Wonka
Computer-aided design (CAD) is the most widely used modeling approach for technical design. The typical starting point in these designs is 2D sketches which can later be extruded and combined to obtain complex three-dimensional assemblies. Such sketches are typically composed of parametric primitives, such as points, lines, and circular arcs, augmented with geometric constraints linking the primitives, such as coincidence, parallelism, or orthogonality. Sketches can be represented as graphs, with the primitives as nodes and the constraints as edges. Training a model to automatically generate CAD sketches can enable several novel workflows, but is challenging due to the complexity of the graphs and the heterogeneity of the primitives and constraints. In particular, each type of primitive and constraint may require a record of different size and parameter types.We propose SketchGen as a generative model based on a transformer architecture to address the heterogeneity problem by carefully designing a sequential language for the primitives and constraints that allows distinguishing between different primitive or constraint types and their parameters, while encouraging our model to re-use information across related parameters, encoding shared structure. A particular highlight of our work is the ability to produce primitives linked via constraints that enables the final output to be further regularized via a constraint solver. We evaluate our model by demonstrating constraint prediction for given sets of primitives and full sketch generation from scratch, showing that our approach significantly out performs the state-of-the-art in CAD sketch generation.
****
CLDA: Contrastive Learning for Semi-Supervised Domain Adaptation

Ankit Singh
Unsupervised Domain Adaptation (UDA) aims to align the labeled source distribution with the unlabeled target distribution to obtain domain invariant predictive models. However, the application of well-known UDA approaches does not generalize well in Semi-Supervised Domain Adaptation (SSDA) scenarios where few labeled samples from the target domain are available.This paper proposes a simple Contrastive Learning framework for semi-supervised Domain Adaptation (CLDA) that attempts to bridge the intra-domain gap between the labeled and unlabeled target distributions and the inter-domain gap between source and unlabeled target distribution in SSDA. We suggest employing class-wise contrastive learning to reduce the inter-domain gap and instance-level contrastive alignment between the original(input image) and strongly augmented unlabeled target images to minimize the intra-domain discrepancy. We have empirically shown that both of these modules complement each other to achieve superior performance. Experiments on three well-known domain adaptation benchmark datasets, namely DomainNet, Office-Home, and Office31, demonstrate the effectiveness of our approach. CLDA achieves state-of-the-art results on all the above datasets.
****
Differentially Private n-gram Extraction
Kunho Kim, Sivakanth Gopi, Janardhan Kulkarni, Sergey Yekhanin
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Capturing implicit hierarchical structure in 3D biomedical images with self-supervised hyperbolic representations
Joy Hsu, Jeffrey Gu, Gong Wu, Wah Chiu, Serena Yeung
We consider the task of representation learning for unsupervised segmentation of 3D voxel-grid biomedical images. We show that models that capture implicit hierarchical relationships between subvolumes are better suited for this task. To that end, we consider encoder-decoder architectures with a hyperbolic latent space, to explicitly capture hierarchical relationships present in subvolumes of the data. We propose utilizing a 3D hyperbolic variational autoencoder with a novel gyroplane convolutional layer to map from the embedding space back to 3D images. To capture these relationships, we introduce an essential self-supervised loss---in addition to the standard VAE loss---which infers approximate hierarchies and encourages implicitly related subvolumes to be mapped closer in the embedding space. We present experiments on synthetic datasets along with a dataset from the medical domain to validate our hypothesis.
****
Noisy Recurrent Neural Networks
Soon Hoe Lim, N. Benjamin Erichson, Liam Hodgkinson, Michael W. Mahoney
We provide a general framework for studying recurrent neural networks (RNNs) trained by injecting noise into hidden states. Specifically, we consider RNNs that can be viewed as discretizations of stochastic differential equations driven by input data. This framework allows us to study the implicit regularization effect of general noise injection schemes by deriving an approximate explicit regularizer in the small noise regime. We find that, under reasonable assumptions, this implicit regularization promotes flatter minima; it biases towards models with more stable dynamics; and, in classification tasks, it favors models with larger classification margin. Sufficient conditions for global stability are obtained, highlighting the phenomenon of stochastic stabilization, where noise injection can improve stability during training. Our theory is supported by empirical results which demonstrate that the RNNs have improved robustness with respect to various input perturbations.
****
Matrix encoding networks for neural combinatorial optimization
Yeong-Dae Kwon, Jinho Choo, Iljoo Yoon, Minah Park, Duwon Park, Youngjune Gwon
Machine Learning (ML) can help solve combinatorial optimization (CO) problems be

tter. A popular approach is to use a neural net to compute on the parameters of a given CO problem and extract useful information that guides the search for good solutions. Many CO problems of practical importance can be specified in a matrix form of parameters quantifying the relationship between two groups of items. There is currently no neural net model, however, that takes in such matrix-style relationship data as an input. Consequently, these types of CO problems have been out of reach for ML engineers. In this paper, we introduce Matrix Encoding Network (MatNet) and show how conveniently it takes in and processes parameters of such complex CO problems. Using an end-to-end model based on MatNet, we solve asymmetric traveling salesman (ATSP) and flexible flow shop (FFSP) problems as the earliest neural approach. In particular, for a class of FFSP we have tested MatNet on, we demonstrate a far superior empirical performance to any methods (neural or not) known to date.
****
When Is Unsupervised Disentanglement Possible?
Daniella Horan, Eitan Richardson, Yair Weiss
A common assumption in many domains is that high dimensional data are a smooth nonlinear function of a small number of independent factors. When is it possible to recover the factors from unlabeled data? In the context of deep models this problem is called “disentanglement” and was recently shown to be impossible without additional strong assumptions [17, 19]. In this paper, we show that the assumption of local isometry together with non-Gaussianity of the factors, is sufficient to provably recover disentangled representations from data. We leverage recent advances in deep generative models to construct manifolds of highly realistic images for which the ground truth latent representation is known, and test whether modern and classical methods succeed in recovering the latent factors. For many different manifolds, we find that a spectral method that explicitly optimizes local isometry and non-Gaussianity consistently finds the correct latent factors, while baseline deep autoencoders do not. We propose how to encourage deep autoencoders to find encodings that satisfy local isometry and show that this helps them discover disentangled representations. Overall, our results suggest that in some realistic settings, unsupervised disentanglement is provably possible, without any domain-specific assumptions.
****
Continuous Latent Process Flows
Ruizhi Deng, Marcus A. Brubaker, Greg Mori, Andreas Lehrmann
Partial observations of continuous time-series dynamics at arbitrary time stamps exist in many disciplines. Fitting this type of data using statistical models with continuous dynamics is not only promising at an intuitive level but also has practical benefits, including the ability to generate continuous trajectories and to perform inference on previously unseen time stamps. Despite exciting progress in this area, the existing models still face challenges in terms of their representational power and the quality of their variational approximations. We tackle these challenges with continuous latent process flows (CLPF), a principled architecture decoding continuous latent processes into continuous observable processes using a time-dependent normalizing flow driven by a stochastic differential equation. To optimize our model using maximum likelihood, we propose a novel piecewise construction of a variational posterior process and derive the corresponding variational lower bound using trajectory re-weighting. Our ablation studies demonstrate the effectiveness of our contributions in various inference tasks on irregular time grids. Comparisons to state-of-the-art baselines show our model's favourable performance on both synthetic and real-world time-series data.
****
Perturbation-based Regret Analysis of Predictive Control in Linear Time Varying Systems
Yiheng Lin, Yang Hu, Guanya Shi, Haoyuan Sun, Guannan Qu, 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.

****
Dataset Distillation with Infinitely Wide Convolutional Networks
Timothy Nguyen, Roman Novak, Lechao Xiao, Jaehoon Lee
The effectiveness of machine learning algorithms arises from being able to extract useful features from large amounts of data. As model and dataset sizes increase, dataset distillation methods that compress large datasets into significantly smaller yet highly performant ones will become valuable in terms of training efficiency and useful feature extraction. To that end, we apply a novel distributed kernel-based meta-learning framework to achieve state-of-the-art results for dataset distillation using infinitely wide convolutional neural networks. For instance, using only 10 datapoints (0.02% of original dataset), we obtain over 65% test accuracy on CIFAR-10 image classification task, a dramatic improvement over the previous best test accuracy of 40%. Our state-of-the-art results extend across many other settings for MNIST, Fashion-MNIST, CIFAR-10, CIFAR-100, and SVHN. Furthermore, we perform some preliminary analyses of our distilled datasets to shed light on how they differ from naturally occurring data.
****
SPANN: Highly-efficient Billion-scale Approximate Nearest Neighborhood Search
Qi Chen, Bing Zhao, Haidong Wang, Mingqin Li, Chuanjie Liu, Zengzhong Li, Mao Yang, Jingdong Wang
The in-memory algorithms for approximate nearest neighbor search (ANNS) have achieved great success for fast high-recall search, but are extremely expensive when handling very large scale database. Thus, there is an increasing request for the hybrid ANNS solutions with small memory and inexpensive solid-state drive (SSD). In this paper, we present a simple but efficient memory-disk hybrid indexing and search system, named SPANN, that follows the inverted index methodology. It stores the centroid points of the posting lists in the memory and the large posting lists in the disk. We guarantee both disk-access efficiency (low latency) and high recall by effectively reducing the disk-access number and retrieving high-quality posting lists. In the index-building stage, we adopt a hierarchical balanced clustering algorithm to balance the length of posting lists and augment the posting list by adding the points in the closure of the corresponding clusters. In the search stage, we use a query-aware scheme to dynamically prune the access of unnecessary posting lists. Experiment results demonstrate that SPANN is 2X faster than the state-of-the-art ANNS solution DiskANN to reach the same recall quality 90% with same memory cost in three billion-scale datasets. It can reach 90% recall@1 and recall@10 in just around one millisecond with only about 10% of original memory cost. Code is available at: https://github.com/microsoft/SPTAG.
****
Distilling Object Detectors with Feature Richness
Du Zhixing, Rui Zhang, Ming Chang, xishan zhang, Shaoli Liu, Tianshi Chen, Yunji Chen
In recent years, large-scale deep models have achieved great success, but the huge computational complexity and massive storage requirements make it a great challenge to deploy them in resource-limited devices. As a model compression and acceleration method, knowledge distillation effectively improves the performance of small models by transferring the dark knowledge from the teacher detector. However, most of the existing distillation-based detection methods mainly imitating features near bounding boxes, which suffer from two limitations. First, they ignore the beneficial features outside the bounding boxes. Second, these methods imitate some features which are mistakenly regarded as the background by the teacher detector. To address the above issues, we propose a novel Feature-Richness Score (FRS) method to choose important features that improve generalized detectability during distilling. The proposed method effectively retrieves the important features outside the bounding boxes and removes the detrimental features within the bounding boxes. Extensive experiments show that our methods achieve excellent performance on both anchor-based and anchor-free detectors. For example, RetinaNet with ResNet-50 achieves 39.7% in mAP on the COCO2017 dataset, which even surpasses the ResNet-101 based teacher detector 38.9% by 0.8%. Our implementation

is available at https://github.com/duzhixing/FRS.
****
Analysis of one-hidden-layer neural networks via the resolvent method
Vanessa Piccolo, Dominik Schröder
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Grounding Spatio-Temporal Language with Transformers
Tristan Karch, Laetitia Teodorescu, Katja Hofmann, Clément Moulin-Frier, Pierre-Yves Oudeyer
Language is an interface to the outside world. In order for embodied agents to use it, language must be grounded in other, sensorimotor modalities. While there is an extended literature studying how machines can learn grounded language, the topic of how to learn spatio-temporal linguistic concepts is still largely uncharted. To make progress in this direction, we here introduce a novel spatio-temporal language grounding task where the goal is to learn the meaning of spatio-temporal descriptions of behavioral traces of an embodied agent. This is achieved by training a truth function that predicts if a description matches a given history of observations. The descriptions involve time-extended predicates in past and present tense as well as spatio-temporal references to objects in the scene. To study the role of architectural biases in this task, we train several models including multimodal Transformer architectures; the latter implement different attention computations between words and objects across space and time. We test models on two classes of generalization: 1) generalization to new sentences, 2) generalization to grammar primitives. We observe that maintaining object identity in the attention computation of our Transformers is instrumental to achieving good performance on generalization overall, and that summarizing object traces in a single token has little influence on performance. We then discuss how this opens new perspectives for language-guided autonomous embodied agents.
****
Learning where to learn: Gradient sparsity in meta and continual learning
Johannes von Oswald, Dominic Zhao, Seijin Kobayashi, Simon Schug, Massimo Caccia, Nicolas Zucchet, João Sacramento
Finding neural network weights that generalize well from small datasets is difficult. A promising approach is to learn a weight initialization such that a small number of weight changes results in low generalization error. We show that this form of meta-learning can be improved by letting the learning algorithm decide which weights to change, i.e., by learning where to learn. We find that patterned sparsity emerges from this process, with the pattern of sparsity varying on a problem-by-problem basis. This selective sparsity results in better generalization and less interference in a range of few-shot and continual learning problems. Moreover, we find that sparse learning also emerges in a more expressive model where learning rates are meta-learned. Our results shed light on an ongoing debate on whether meta-learning can discover adaptable features and suggest that learning by sparse gradient descent is a powerful inductive bias for meta-learning systems.
****
Domain Invariant Representation Learning with Domain Density Transformations
A. Tuan Nguyen, Toan Tran, Yarin Gal, Atilim Gunes Baydin
Domain generalization refers to the problem where we aim to train a model on data from a set of source domains so that the model can generalize to unseen target domains. Naively training a model on the aggregate set of data (pooled from all source domains) has been shown to perform suboptimally, since the information learned by that model might be domain-specific and generalize imperfectly to target domains. To tackle this problem, a predominant domain generalization approach is to learn some domain-invariant information for the prediction task, aiming at a good generalization across domains. In this paper, we propose a theoretically grounded method to learn a domain-invariant representation by enforcing the r

epresentation network to be invariant under all transformation functions among domains. We next introduce the use of generative adversarial networks to learn such domain transformations in a possible implementation of our method in practice. We demonstrate the effectiveness of our method on several widely used datasets for the domain generalization problem, on all of which we achieve competitive results with state-of-the-art models.
****
PlayVirtual: Augmenting Cycle-Consistent Virtual Trajectories for Reinforcement Learning
Tao Yu, Cuiling Lan, Wenjun Zeng, Mingxiao Feng, Zhizheng Zhang, Zhibo Chen
Learning good feature representations is important for deep reinforcement learning (RL). However, with limited experience, RL often suffers from data inefficiency for training. For un-experienced or less-experienced trajectories (i.e., state-action sequences), the lack of data limits the use of them for better feature learning. In this work, we propose a novel method, dubbed PlayVirtual, which augments cycle-consistent virtual trajectories to enhance the data efficiency for RL feature representation learning. Specifically, PlayVirtual predicts future states in a latent space based on the current state and action by a dynamics model and then predicts the previous states by a backward dynamics model, which forms a trajectory cycle. Based on this, we augment the actions to generate a large amount of virtual state-action trajectories. Being free of groudtruth state supervision, we enforce a trajectory to meet the cycle consistency constraint, which can significantly enhance the data efficiency. We validate the effectiveness of our designs on the Atari and DeepMind Control Suite benchmarks. Our method achieves the state-of-the-art performance on both benchmarks. Our code is available at https://github.com/microsoft/Playvirtual.
****
Efficient Equivariant Network
Lingshen He, Yuxuan Chen, zhengyang shen, Yiming Dong, Yisen Wang, Zhouchen 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-authors prior to requesting a name change in the electronic proceedings.
****
Unifying Gradient Estimators for Meta-Reinforcement Learning via Off-Policy Evaluation
Yunhao Tang, Tadashi Kozuno, Mark Rowland, Remi Munos, Michal Valko
Model-agnostic meta-reinforcement learning requires estimating the Hessian matrix of value functions. This is challenging from an implementation perspective, as repeatedly differentiating policy gradient estimates may lead to biased Hessian estimates. In this work, we provide a unifying framework for estimating higher-order derivatives of value functions, based on off-policy evaluation. Our framework interprets a number of prior approaches as special cases and elucidates the bias and variance trade-off of Hessian estimates. This framework also opens the door to a new family of estimates, which can be easily implemented with auto-differentiation libraries, and lead to performance gains in practice.
****
Even your Teacher Needs Guidance: Ground-Truth Targets Dampen Regularization Imposed by Self-Distillation
Kenneth Borup, Lars N Andersen
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Compressing Neural Networks: Towards Determining the Optimal Layer-wise Decomposition
Lucas Liebenwein, Alaa Maalouf, Dan Feldman, Daniela Rus
We present a novel global compression framework for deep neural networks that automatically analyzes each layer to identify the optimal per-layer compression ra

tio, while simultaneously achieving the desired overall compression. Our algorit
hm hinges on the idea of compressing each convolutional (or fully-connected) lay
er by slicing its channels into multiple groups and decomposing each group via l
ow-rank decomposition. At the core of our algorithm is the derivation of layer-w
ise error bounds from the Eckart–Young–Mirsky theorem. We then leverage these bo
unds to frame the compression problem as an optimization problem where we wish t
o minimize the maximum compression error across layers and propose an efficient
algorithm towards a solution. Our experiments indicate that our method outperfor
ms existing low-rank compression approaches across a wide range of networks and
data sets. We believe that our results open up new avenues for future research i
nto the global performance-size trade-offs of modern neural networks.
****
Equilibrium and non-Equilibrium regimes in the learning of Restricted Boltzmann
Machines
Aurélien Decelle, Cyril Furtlehner, Beatriz Seoane
Requests for name changes in the electronic proceedings will be accepted with 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.
****
Imitation with Neural Density Models
Kuno Kim, Akshat Jindal, Yang Song, Jiaming Song, Yanan Sui, Stefano Ermon
We propose a new framework for Imitation Learning (IL) via density estimation of
 the expert's occupancy measure followed by Maximum Occupancy Entropy Reinforcem
ent Learning (RL) using the density as a reward. Our approach maximizes a non-ad
versarial model-free RL objective that provably lower bounds reverse Kullback–Le
ibler divergence between occupancy measures of the expert and imitator. We prese
nt a practical IL algorithm, Neural Density Imitation (NDI), which obtains state
-of-the-art demonstration efficiency on benchmark control tasks.
****
Accurate Point Cloud Registration with Robust Optimal Transport
Zhengyang Shen, Jean Feydy, Peirong Liu, Ariel H Curiale, Ruben San Jose Estepar
, Raul San Jose Estepar, Marc Niethammer
This work investigates the use of robust optimal transport (OT) for shape matchi
ng. Specifically, we show that recent OT solvers improve both optimization-based
 and deep learning methods for point cloud registration, boosting accuracy at an
 affordable computational cost. This manuscript starts with a practical overview
 of modern OT theory. We then provide solutions to the main difficulties in usin
g this framework for shape matching. Finally, we showcase the performance of tra
nsport-enhanced registration models on a wide range of challenging tasks: rigid
registration for partial shapes; scene flow estimation on the Kitti dataset; and
 nonparametric registration of lung vascular trees between inspiration and expir
ation. Our OT-based methods achieve state-of-the-art results on Kitti and for th
e challenging lung registration task, both in terms of accuracy and scalability.
 We also release PVT1010, a new public dataset of 1,010 pairs of lung vascular t
rees with densely sampled points. This dataset provides a challenging use case f
or point cloud registration algorithms with highly complex shapes and deformatio
ns. Our work demonstrates that robust OT enables fast pre-alignment and fine-tun
ing for a wide range of registration models, thereby providing a new key method
for the computer vision toolbox. Our code and dataset are available online at: h
ttps://github.com/uncbiag/robot.
****
Simple steps are all you need: Frank-Wolfe and generalized self-concordant funct
ions
Alejandro Carderera, Mathieu Besançon, Sebastian Pokutta
Requests for name changes in the electronic proceedings will be accepted with 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 Data Augmentation for Generalization in Reinforcement Learning
Roberta Raileanu, Maxwell Goldstein, Denis Yarats, Ilya Kostrikov, Rob Fergus
Deep reinforcement learning (RL) agents often fail to generalize beyond their training environments. To alleviate this problem, recent work has proposed the use of data augmentation. However, different tasks tend to benefit from different types of augmentations and selecting the right one typically requires expert knowledge. In this paper, we introduce three approaches for automatically finding an effective augmentation for any RL task. These are combined with two novel regularization terms for the policy and value function, required to make the use of data augmentation theoretically sound for actor-critic algorithms. Our method achieves a new state-of-the-art on the Procgen benchmark and outperforms popular RL algorithms on DeepMind Control tasks with distractors. In addition, our agent learns policies and representations which are more robust to changes in the environment that are irrelevant for solving the task, such as the background.
****
Blending Anti-Aliasing into Vision Transformer
Shengju Qian, Hao Shao, Yi Zhu, Mu Li, Jiaya Jia
The transformer architectures, based on self-attention mechanism and convolution-free design, recently found superior performance and booming applications in computer vision. However, the discontinuous patch-wise tokenization process implicitly introduces jagged artifacts into attention maps, arising the traditional problem of aliasing for vision transformers. Aliasing effect occurs when discrete patterns are used to produce high frequency or continuous information, resulting in the indistinguishable distortions. Recent researches have found that modern convolution networks still suffer from this phenomenon. In this work, we analyze the uncharted problem of aliasing in vision transformer and explore to incorporate anti-aliasing properties. Specifically, we propose a plug-and-play Aliasing-Reduction Module (ARM) to alleviate the aforementioned issue. We investigate the effectiveness and generalization of the proposed method across multiple tasks and various vision transformer families. This lightweight design consistently attains a clear boost over several famous structures. Furthermore, our module also improves data efficiency and robustness of vision transformers.
****
A Trainable Spectral-Spatial Sparse Coding Model for Hyperspectral Image Restoration
Theo Bodrito, Alexandre Zouaoui, Jocelyn Chanussot, Julien Mairal
Hyperspectral imaging offers new perspectives for diverse applications, ranging from the monitoring of the environment using airborne or satellite remote sensing, precision farming, food safety, planetary exploration, or astrophysics. Unfortunately, the spectral diversity of information comes at the expense of various sources of degradation, and the lack of accurate ground-truth "clean" hyperspectral signals acquired on the spot makes restoration tasks challenging. In particular, training deep neural networks for restoration is difficult, in contrast to traditional RGB imaging problems where deep models tend to shine. In this paper, we advocate instead for a hybrid approach based on sparse coding principles that retain the interpretability of classical techniques encoding domain knowledge with handcrafted image priors, while allowing to train model parameters end-to-end without massive amounts of data. We show on various denoising benchmarks that our method is computationally efficient and significantly outperforms the state of the art.
****
Posterior Collapse and Latent Variable Non-identifiability
Yixin Wang, David Blei, John P. Cunningham
Variational autoencoders model high-dimensional data by positinglow-dimensional latent variables that are mapped through a flexibledistribution parametrized by a neural network. Unfortunately,variational autoencoders often suffer from posterior collapse: theposterior of the latent variables is equal to its prior, rendering thevariational autoencoder useless as a means to produce meaningfulrepresentations. Existing approaches to posterior collapse oftenattribute it to the use of neural networks or optimization issues dueto variational approximation. In th

is paper, we consider posteriorcollapse as a problem of latent variable non-iden
tifiability. We provethat the posterior collapses if and only if the latent vari
ables arenon-identifiable in the generative model. This fact implies thatposteri
or collapse is not a phenomenon specific to the use of flexibledistributions or
approximate inference. Rather, it can occur inclassical probabilistic models eve
n with exact inference, which wealso demonstrate. Based on these results, we pro
pose a class oflatent-identifiable variational autoencoders, deep generative mod
elswhich enforce identifiability without sacrificing flexibility. Thismodel clas
s resolves the problem of latent variablenon-identifiability by leveraging bijec
tive Brenier maps andparameterizing them with input convex neural networks, with
out specialvariational inference objectives or optimization tricks. Acrosssynthe
tic and real datasets, latent-identifiable variationalautoencoders outperform ex
isting methods in mitigating posteriorcollapse and providing meaningful represen
tations of the data.
****
The Benefits of Implicit Regularization from SGD in Least Squares Problems
Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, Dean P. Foster, Sham Ka
kade
Stochastic gradient descent (SGD) exhibits strong algorithmic regularization eff
ects in practice, which has been hypothesized to play an important role in the g
eneralization of modern machine learning approaches. In this work, we seek to un
derstand these issues in the simpler setting of linear regression (including bot
h underparameterized and overparameterized regimes), where our goal is to make s
harp instance-based comparisons of the implicit regularization afforded by (unre
gularized) average SGD with the explicit regularization of ridge regression. For
 a broad class of least squares problem instances (that are natural in high-dime
nsional settings), we show: (1) for every problem instance and for every ridge p
arameter, (unregularized) SGD, when provided with \emph{logarithmically} more sa
mples than that provided to the ridge algorithm, generalizes no worse than the r
idge solution (provided SGD uses a tuned constant stepsize); (2) conversely, the
re exist instances (in this wide problem class) where optimally-tuned ridge regr
ession requires \emph{quadratically} more samples than SGD in order to have the
same generalization performance. Taken together, our results show that, up to th
e logarithmic factors, the generalization performance of SGD is always no worse
than that of ridge regression in a wide range of overparameterized problems, and
, in fact, could be much better for some problem instances. More generally, our
results show how algorithmic regularization has important consequences even in s
impler (overparameterized) convex settings.
****
Generalization of Model-Agnostic Meta-Learning Algorithms: Recurring and Unseen
Tasks
Alireza Fallah, Aryan Mokhtari, Asuman Ozdaglar
Requests for name changes in the electronic proceedings will be accepted with 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.
****
Factored Policy Gradients: Leveraging Structure for Efficient Learning in MOMDPs
Thomas Spooner, Nelson Vadori, Sumitra Ganesh
Policy gradient methods can solve complex tasks but often fail when the dimensio
nality of the action-space or objective multiplicity grow very large. This occur
s, in part, because the variance on score-based gradient estimators scales quadr
atically. In this paper, we address this problem through a factor baseline which
 exploits independence structure encoded in a novel action-target influence netw
ork. Factored policy gradients (FPGs), which follow, provide a common framework
for analysing key state-of-the-art algorithms, are shown to generalise tradition
al policy gradients, and yield a principled way of incorporating prior knowledge
 of a problem domain's generative processes. We provide an analysis of the propo
sed estimator and identify the conditions under which variance is reduced. The a
lgorithmic aspects of FPGs are discussed, including optimal policy factorisation

, as characterised by minimum biclique coverings, and the implications for the bias variance trade-off of incorrectly specifying the network. Finally, we demonstrate the performance advantages of our algorithm on large-scale bandit and traffic intersection problems, providing a novel contribution to the latter in the form of a spatial approximation.
****

MarioNette: Self-Supervised Sprite Learning

Dmitriy Smirnov, MICHAEL GHARBI, Matthew Fisher, Vitor Guizilini, Alexei Efros, Justin M. Solomon

Artists and video game designers often construct 2D animations using libraries of sprites---textured patches of objects and characters. We propose a deep learning approach that decomposes sprite-based video animations into a disentangled representation of recurring graphic elements in a self-supervised manner. By jointly learning a dictionary of possibly transparent patches and training a network that places them onto a canvas, we deconstruct sprite-based content into a sparse, consistent, and explicit representation that can be easily used in downstream tasks, like editing or analysis. Our framework offers a promising approach for discovering recurring visual patterns in image collections without supervision.
****

RLlib Flow: Distributed Reinforcement Learning is a Dataflow Problem

Eric Liang, Zhanghao Wu, Michael Luo, Sven Mika, Joseph E. Gonzalez, Ion Stoica

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

Improve Agents without Retraining: Parallel Tree Search with Off-Policy Correction

Gal Dalal, Assaf Hallak, Steven Dalton, iuri frosio, Shie Mannor, Gal Chechik

Tree Search (TS) is crucial to some of the most influential successes in reinforcement learning. Here, we tackle two major challenges with TS that limit its usability: \textit{distribution shift} and \textit{scalability}. We first discover and analyze a counter-intuitive phenomenon: action selection through TS and a pre-trained value function often leads to lower performance compared to the original pre-trained agent, even when having access to the exact state and reward in future steps. We show this is due to a distribution shift to areas where value estimates are highly inaccurate and analyze this effect using Extreme Value theory. To overcome this problem, we introduce a novel off-policy correction term that accounts for the mismatch between the pre-trained value and its corresponding TS policy by penalizing under-sampled trajectories. We prove that our correction eliminates the above mismatch and bound the probability of sub-optimal action selection. Our correction significantly improves pre-trained Rainbow agents without any further training, often more than doubling their scores on Atari games. Next, we address the scalability issue given by the computational complexity of exhaustive TS that scales exponentially with the tree depth. We introduce Batch-BFS: a GPU breadth-first search that advances all nodes in each depth of the tree simultaneously. Batch-BFS reduces runtime by two orders of magnitude and, beyond inference, enables also training with TS of depths that were not feasible before. We train DQN agents from scratch using TS and show improvement in several Atari games compared to both the original DQN and the more advanced Rainbow. We will share the code upon publication.
****

Redesigning the Transformer Architecture with Insights from Multi-particle Dynamical Systems

Subhabrata Dutta, Tanya Gautam, Soumen Chakrabarti, Tanmoy Chakraborty

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

Exploring Architectural Ingredients of Adversarially Robust Deep Neural Networks
Hanxun Huang, Yisen Wang, Sarah Erfani, Quanquan Gu, James Bailey, Xingjun Ma
Deep neural networks (DNNs) are known to be vulnerable to adversarial attacks. A range of defense methods have been proposed to train adversarially robust DNNs, among which adversarial training has demonstrated promising results. However, despite preliminary understandings developed for adversarial training, it is still not clear, from the architectural perspective, what configurations can lead to more robust DNNs. In this paper, we address this gap via a comprehensive investigation on the impact of network width and depth on the robustness of adversarially trained DNNs. Specifically, we make the following key observations: 1) more parameters (higher model capacity) does not necessarily help adversarial robustness; 2) reducing capacity at the last stage (the last group of blocks) of the network can actually improve adversarial robustness; and 3) under the same parameter budget, there exists an optimal architectural configuration for adversarial robustness. We also provide a theoretical analysis explaning why such network configuration can help robustness. These architectural insights can help design adversarially robust DNNs.
****
Center Smoothing: Certified Robustness for Networks with Structured Outputs
Aounon Kumar, Tom Goldstein
The study of provable adversarial robustness has mostly been limited to classification tasks and models with one-dimensional real-valued outputs. We extend the scope of certifiable robustness to problems with more general and structured outputs like sets, images, language, etc. We model the output space as a metric space under a distance/similarity function, such as intersection-over-union, perceptual similarity, total variation distance, etc. Such models are used in many machine learning problems like image segmentation, object detection, generative models, image/audio-to-text systems, etc. Based on a robustness technique called randomized smoothing, our center smoothing procedure can produce models with the guarantee that the change in the output, as measured by the distance metric, remains small for any norm-bounded adversarial perturbation of the input. We apply our method to create certifiably robust models with disparate output spaces -- from sets to images -- and show that it yields meaningful certificates without significantly degrading the performance of the base model.
****
Breaking the Linear Iteration Cost Barrier for Some Well-known Conditional Gradient Methods Using MaxIP Data-structures
Zhaozhuo Xu, Zhao Song, Anshumali Shrivastava
Conditional gradient methods (CGM) are widely used in modern machine learning. CGM's overall running time usually consists of two parts: the number of iterations and the cost of each iteration. Most efforts focus on reducing the number of iterations as a means to reduce the overall running time. In this work, we focus on improving the per iteration cost of CGM. The bottleneck step in most CGM is maximum inner product search (MaxIP), which requires a linear scan over the parameters. In practice, approximate MaxIP data-structures are found to be helpful heuristics. However, theoretically, nothing is known about the combination of approximate MaxIP data-structures and CGM. In this work, we answer this question positively by providing a formal framework to combine the locality sensitive hashing type approximate MaxIP data-structures with CGM algorithms. As a result, we show the first algorithm, where the cost per iteration is sublinear in the number of parameters, for many fundamental optimization algorithms, e.g., Frank-Wolfe, Herding algorithm, and policy gradient.
****
Neural Regression, Representational Similarity, Model Zoology & Neural Taskonomy at Scale in Rodent Visual Cortex
Colin Conwell, David Mayo, Andrei Barbu, Michael Buice, George Alvarez, Boris Katz
How well do deep neural networks fare as models of mouse visual cortex? A majority of research to date suggests results far more mixed than those produced in the modeling of primate visual cortex. Here, we perform a large-scale benchmarking

of dozens of deep neural network models in mouse visual cortex with both representational similarity analysis and neural regression. Using the Allen Brain Observatory's 2-photon calcium-imaging dataset of activity in over 6,000 reliable rodent visual cortical neurons recorded in response to natural scenes, we replicate previous findings and resolve previous discrepancies, ultimately demonstrating that modern neural networks can in fact be used to explain activity in the mouse visual cortex to a more reasonable degree than previously suggested. Using our benchmark as an atlas, we offer preliminary answers to overarching questions about levels of analysis (e.g. do models that better predict the representations of individual neurons also predict representational similarity across neural populations?); questions about the properties of models that best predict the visual system overall (e.g. is convolution or category-supervision necessary to better predict neural activity?); and questions about the mapping between biological and artificial representations (e.g. does the information processing hierarchy in deep nets match the anatomical hierarchy of mouse visual cortex?). Along the way, we catalogue a number of models (including vision transformers, MLP-Mixers, normalization free networks, Taskonomy encoders and self-supervised models) outside the traditional circuit of convolutional object recognition. Taken together, our results provide a reference point for future ventures in the deep neural network modeling of mouse visual cortex, hinting at novel combinations of mapping method, architecture, and task to more fully characterize the computational motifs of visual representation in a species so central to neuroscience, but with a perceptual physiology and ecology markedly different from the ones we study in primates.

****

A Topological Perspective on Causal Inference

Duligur Ibeling, Thomas Icard

This paper presents a topological learning-theoretic perspective on causal inference by introducing a series of topologies defined on general spaces of structural causal models (SCMs). As an illustration of the framework we prove a topological causal hierarchy theorem, showing that substantive assumption-free causal inference is possible only in a meager set of SCMs. Thanks to a known correspondence between open sets in the weak topology and statistically verifiable hypotheses, our results show that inductive assumptions sufficient to license valid causal inferences are statistically unverifiable in principle. Similar to no-free-lunch theorems for statistical inference, the present results clarify the inevitability of substantial assumptions for causal inference. An additional benefit of our topological approach is that it easily accommodates SCMs with infinitely many variables. We finally suggest that our framework may be helpful for the positive project of exploring and assessing alternative causal-inductive assumptions.

****

Parameter Inference with Bifurcation Diagrams

Gregory Szep, Neil Dalchau, Attila Csikász-Nagy

Estimation of parameters in differential equation models can be achieved by applying learning algorithms to quantitative time-series data. However, sometimes it is only possible to measure qualitative changes of a system in response to a controlled condition. In dynamical systems theory, such change points are known as bifurcations and lie on a function of the controlled condition called the bifurcation diagram. In this work, we propose a gradient-based approach for inferring the parameters of differential equations that produce a user-specified bifurcation diagram. The cost function contains an error term that is minimal when the model bifurcations match the specified targets and a bifurcation measure which has gradients that push optimisers towards bifurcating parameter regimes. The gradients can be computed without the need to differentiate through the operations of the solver that was used to compute the diagram. We demonstrate parameter inference with minimal models which explore the space of saddle-node and pitchfork diagrams and the genetic toggle switch from synthetic biology. Furthermore, the cost landscape allows us to organise models in terms of topological and geometric equivalence.

****

Scalable Thompson Sampling using Sparse Gaussian Process Models
Sattar Vakili, Henry Moss, Artem Artemev, Vincent Dutordoir, Victor Picheny
Thompson Sampling (TS) from Gaussian Process (GP) models is a powerful tool for
the optimization of black-box functions. Although TS enjoys strong theoretical g
uarantees and convincing empirical performance, it incurs a large computational
overhead that scales polynomially with the optimization budget. Recently, scalab
le TS methods based on sparse GP models have been proposed to increase the scope
 of TS, enabling its application to problems that are sufficiently multi-modal,
noisy or combinatorial to require more than a few hundred evaluations to be solv
ed. However, the approximation error introduced by sparse GPs invalidates all ex
isting regret bounds. In this work, we perform a theoretical and empirical analy
sis of scalable TS. We provide theoretical guarantees and show that the drastic
reduction in computational complexity of scalable TS can be enjoyed without loss
 in the regret performance over the standard TS. These conceptual claims are val
idated for practical implementations of scalable TS on synthetic benchmarks and
as part of a real-world high-throughput molecular design task.
****
Robust Counterfactual Explanations on Graph Neural Networks
Mohit Bajaj, Lingyang Chu, Zi Yu Xue, Jian Pei, Lanjun Wang, Peter Cho-Ho Lam, Y
ong Zhang
Massive deployment of Graph Neural Networks (GNNs) in high-stake applications ge
nerates a strong demand for explanations that are robust to noise and align well
 with human intuition. Most existing methods generate explanations by identifyin
g a subgraph of an input graph that has a strong correlation with the prediction
. These explanations are not robust to noise because independently optimizing th
e correlation for a single input can easily overfit noise. Moreover, they are no
t counterfactual because removing an identified subgraph from an input graph doe
s not necessarily change the prediction result. In this paper, we propose a nove
l method to generate robust counterfactual explanations on GNNs by explicitly mo
delling the common decision logic of GNNs on similar input graphs. Our explanati
ons are naturally robust to noise because they are produced from the common deci
sion boundaries of a GNN that govern the predictions of many similar input graph
s. The explanations are also counterfactual because removing the set of edges id
entified by an explanation from the input graph changes the prediction significa
ntly. Exhaustive experiments on many public datasets demonstrate the superior pe
rformance of our method.
****
Similarity and Matching of Neural Network Representations
Adrián Csiszárik, Péter K■rösi-Szabó, Ákos Matszangosz, Gergely Papp, Dániel Var
ga
We employ a toolset --- dubbed Dr. Frankenstein --- to analyse the similarity of
 representations in deep neural networks. With this toolset we aim to match the
activations on given layers of two trained neural networks by joining them with
a stitching layer. We demonstrate that the inner representations emerging in dee
p convolutional neural networks with the same architecture but different initial
isations can be matched with a surprisingly high degree of accuracy even with a
single, affine stitching layer. We choose the stitching layer from several possi
ble classes of linear transformations and investigate their performance and prop
erties. The task of matching representations is closely related to notions of si
milarity. Using this toolset we also provide a novel viewpoint on the current li
ne of research regarding similarity indices of neural network representations: t
he perspective of the performance on a task.
****
DOCTOR: A Simple Method for Detecting Misclassification Errors
Federica Granese, Marco Romanelli, Daniele Gorla, Catuscia Palamidessi, Pablo Pi
antanida
Deep neural networks (DNNs) have shown to perform very well on large scale objec
t recognition problems and lead to widespread use for real-world applications, i
ncluding situations where DNN are implemented as “black boxes”. A promising app
roach to secure their use is to accept decisions that are likely to be correct w

hile discarding the others. In this work, we propose DOCTOR, a simple method th
at aims to identify whether the prediction of a DNN classifier should (or should
 not) be trusted so that, consequently, it would be possible to accept it or to
reject it. Two scenarios are investigated: Totally Black Box (TBB) where only th
e soft-predictions are available and Partially Black Box (PBB) where gradient-pr
opagation to perform input pre-processing is allowed. Empirically, we show that
DOCTOR outperforms all state-of-the-art methods on various well-known images and
 sentiment analysis datasets. In particular, we observe a reduction of up to 4%
of the false rejection rate (FRR) in the PBB scenario. DOCTOR can be applied to
any pre-trained model, it does not require prior information about the underlyin
g dataset and is as simple as the simplest available methods in the literature.
****
Contrastive Laplacian Eigenmaps
Hao Zhu, Ke Sun, Peter Koniusz
Graph contrastive learning attracts/disperses node representations for similar/d
issimilar node pairs under some notion of similarity. It may be combined with a
low-dimensional embedding of nodes to preserve intrinsic and structural properti
es of a graph. In this paper, we extend the celebrated Laplacian Eigenmaps with
contrastive learning, and call them COntrastive Laplacian EigenmapS (COLES). Sta
rting from a GAN-inspired contrastive formulation, we show that the Jensen-Shann
on divergence underlying many contrastive graph embedding models fails under dis
joint positive and negative distributions, which may naturally emerge during sam
pling in the contrastive setting. In contrast, we demonstrate analytically that
COLES essentially minimizes a surrogate of Wasserstein distance, which is known
to cope well under disjoint distributions. Moreover, we show that the loss of CO
LES belongs to the family of so-called block-contrastive losses, previously show
n to be superior compared to pair-wise losses typically used by contrastive meth
ods. We show on popular benchmarks/backbones that COLES offers favourable accura
cy/scalability compared to DeepWalk, GCN, Graph2Gauss, DGI and GRACE baselines.
****
Machine learning structure preserving brackets for forecasting irreversible proc
esses
Kookjin Lee, Nathaniel Trask, Panos Stinis
Forecasting of time-series data requires imposition of inductive biases to obtai
n predictive extrapolation, and recent works have imposed Hamiltonian/Lagrangian
 form to preserve structure for systems with \emph{reversible} dynamics. In this
 work we present a novel parameterization of dissipative brackets from metriplec
tic dynamical systems appropriate for learning \emph{irreversible} dynamics with
 unknown a priori model form. The process learns generalized Casimirs for energy
 and entropy guaranteed to be conserved and nondecreasing, respectively. Further
more, for the case of added thermal noise, we guarantee exact preservation of a
fluctuation-dissipation theorem, ensuring thermodynamic consistency. We provide
benchmarks for dissipative systems demonstrating learned dynamics are more robus
t and generalize better than either "black-box" or penalty-based approaches.
****
On the Variance of the Fisher Information for Deep Learning
Alexander Soen, Ke Sun
In the realm of deep learning, the Fisher information matrix (FIM) gives novel i
nsights and useful tools to characterize the loss landscape, perform second-orde
r optimization, and build geometric learning theories. The exact FIM is either u
navailable in closed form or too expensive to compute. In practice, it is almost
 always estimated based on empirical samples. We investigate two such estimator
s based on two equivalent representations of the FIM --- both unbiased and consi
stent. Their estimation quality is naturally gauged by their variance given in c
losed form. We analyze how the parametric structure of a deep neural network can
 affect the variance. The meaning of this variance measure and its upper bounds
are then discussed in the context of deep learning.
****
A^2-Net: Learning Attribute-Aware Hash Codes for Large-Scale Fine-Grained Imag
e Retrieval

Xiu-Shen Wei, Yang Shen, Xuhao Sun, Han-Jia Ye, Jian 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.
****
Shape Registration in the Time of Transformers
Giovanni Trappolini, Luca Cosmo, Luca Moschella, Riccardo Marin, Simone Melzi, Emanuele Rodolà
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Brick-by-Brick: Combinatorial Construction with Deep Reinforcement Learning
Hyunsoo Chung, Jungtaek Kim, Boris Knyazev, Jinhwi Lee, Graham W. Taylor, Jaesik Park, Minsu Cho
Discovering a solution in a combinatorial space is prevalent in many real-world problems but it is also challenging due to diverse complex constraints and the vast number of possible combinations. To address such a problem, we introduce a novel formulation, combinatorial construction, which requires a building agent to assemble unit primitives (i.e., LEGO bricks) sequentially -- every connection between two bricks must follow a fixed rule, while no bricks mutually overlap. To construct a target object, we provide incomplete knowledge about the desired target (i.e., 2D images) instead of exact and explicit volumetric information to the agent. This problem requires a comprehensive understanding of partial information and long-term planning to append a brick sequentially, which leads us to employ reinforcement learning. The approach has to consider a variable-sized action space where a large number of invalid actions, which would cause overlap between bricks, exist. To resolve these issues, our model, dubbed Brick-by-Brick, adopts an action validity prediction network that efficiently filters invalid actions for an actor-critic network. We demonstrate that the proposed method successfully learns to construct an unseen object conditioned on a single image or multiple views of a target object.
****
Dissecting the Diffusion Process in Linear Graph Convolutional Networks
Yifei Wang, Yisen Wang, Jiansheng Yang, Zhouchen Lin
Graph Convolutional Networks (GCNs) have attracted more and more attentions in recent years. A typical GCN layer consists of a linear feature propagation step and a nonlinear transformation step. Recent works show that a linear GCN can achieve comparable performance to the original non-linear GCN while being much more computationally efficient. In this paper, we dissect the feature propagation steps of linear GCNs from a perspective of continuous graph diffusion, and analyze why linear GCNs fail to benefit from more propagation steps. Following that, we propose Decoupled Graph Convolution (DGC) that decouples the terminal time and the feature propagation steps, making it more flexible and capable of exploiting a very large number of feature propagation steps. Experiments demonstrate that our proposed DGC improves linear GCNs by a large margin and makes them competitive with many modern variants of non-linear GCNs.
****
Dynamic Grained Encoder for Vision Transformers
Lin Song, Songyang Zhang, Songtao Liu, Zeming Li, Xuming He, Hongbin Sun, Jian Sun, Nanning Zheng
Transformers, the de-facto standard for language modeling, have been recently applied for vision tasks. This paper introduces sparse queries for vision transformers to exploit the intrinsic spatial redundancy of natural images and save computational costs. Specifically, we propose a Dynamic Grained Encoder for vision transformers, which can adaptively assign a suitable number of queries to each spatial region. Thus it achieves a fine-grained representation in discriminative regions while keeping high efficiency. Besides, the dynamic grained encoder is co

mpatible with most vision transformer frameworks. Without bells and whistles, our encoder allows the state-of-the-art vision transformers to reduce computational complexity by 40%-60% while maintaining comparable performance on image classification. Extensive experiments on object detection and segmentation further demonstrate the generalizability of our approach. Code is available at https://github.com/StevenGrove/vtpack.
****
Understanding Negative Samples in Instance Discriminative Self-supervised Representation Learning
Kento Nozawa, Issei Sato
Instance discriminative self-supervised representation learning has been attracted attention thanks to its unsupervised nature and informative feature representation for downstream tasks. In practice, it commonly uses a larger number of negative samples than the number of supervised classes. However, there is an inconsistency in the existing analysis; theoretically, a large number of negative samples degrade classification performance on a downstream supervised task, while empirically, they improve the performance. We provide a novel framework to analyze this empirical result regarding negative samples using the coupon collector's problem. Our bound can implicitly incorporate the supervised loss of the downstream task in the self-supervised loss by increasing the number of negative samples. We confirm that our proposed analysis holds on real-world benchmark datasets.
****
On UMAP's True Loss Function
Sebastian Damrich, Fred A. Hamprecht
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 Pure Exploration via Frank-Wolfe
Po-An Wang, Ruo-Chun Tzeng, Alexandre Proutiere
We study the problem of active pure exploration with fixed confidence in generic stochastic bandit environments. The goal of the learner is to answer a query about the environment with a given level of certainty while minimizing her sampling budget. For this problem, instance-specific lower bounds on the expected sample complexity reveal the optimal proportions of arm draws an Oracle algorithm would apply. These proportions solve an optimization problem whose tractability strongly depends on the structural properties of the environment, but may be instrumental in the design of efficient learning algorithms. We devise Frank-Wolfe-based Sampling (FWS), a simple algorithm whose sample complexity matches the lower bounds for a wide class of pure exploration problems. The algorithm is computationally efficient as, to learn and track the optimal proportion of arm draws, it relies on a single iteration of Frank-Wolfe algorithm applied to the lower-bound optimization problem. We apply FWS to various pure exploration tasks, including best arm identification in unstructured, thresholded, linear, and Lipschitz bandits. Despite its simplicity, FWS is competitive compared to state-of-art algorithms.
****
iFlow: Numerically Invertible Flows for Efficient Lossless Compression via a Uniform Coder
Shifeng Zhang, Ning Kang, Tom Ryder, Zhenguo 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.
****
History Aware Multimodal Transformer for Vision-and-Language Navigation
Shizhe Chen, Pierre-Louis Guhur, Cordelia Schmid, Ivan Laptev
Vision-and-language navigation (VLN) aims to build autonomous visual agents that follow instructions and navigate in real scenes. To remember previously visited

locations and actions taken, most approaches to VLN implement memory using recu rrent states. Instead, we introduce a History Aware Multimodal Transformer (HAMT) to incorporate a long-horizon history into multimodal decision making. HAMT ef ficiently encodes all the past panoramic observations via a hierarchical vision transformer (ViT), which first encodes individual images with ViT, then models s patial relation between images in a panoramic observation and finally takes into account temporal relation between panoramas in the history. It, then, jointly c ombines text, history and current observation to predict the next action. We fir st train HAMT end-to-end using several proxy tasks including single step action prediction and spatial relation prediction, and then use reinforcement learning to further improve the navigation policy. HAMT achieves new state of the art on a broad range of VLN tasks, including VLN with fine-grained instructions (R2R, R xR), high-level instructions (R2R-Last, REVERIE), dialogs (CVDN) as well as long -horizon VLN (R4R, R2R-Back). We demonstrate HAMT to be particularly effective f or navigation tasks with longer trajectories.
****
Meta Two-Sample Testing: Learning Kernels for Testing with Limited Data
Feng Liu, Wenkai Xu, Jie Lu, Danica J. Sutherland
Modern kernel-based two-sample tests have shown great success in distinguishing complex, high-dimensional distributions by learning appropriate kernels (or, as a special case, classifiers). Previous work, however, has assumed that many samp les are observed from both of the distributions being distinguished. In realisti c scenarios with very limited numbers of data samples, it can be challenging to identify a kernel powerful enough to distinguish complex distributions. We addre ss this issue by introducing the problem of meta two-sample testing (M2ST), whic h aims to exploit (abundant) auxiliary data on related tasks to find an algorith m that can quickly identify a powerful test on new target tasks. We propose two specific algorithms for this task: a generic scheme which improves over baseline s, and a more tailored approach which performs even better. We provide both theo retical justification and empirical evidence that our proposed meta-testing sche mes outperform learning kernel-based tests directly from scarce observations, an d identify when such schemes will be successful.
****
Process for Adapting Language Models to Society (PALMS) with Values-Targeted Dat asets
Irene Solaiman, Christy Dennison
Language models can generate harmful and biased outputs and exhibit undesirable behavior according to a given cultural context. We propose a Process for Adaptin g Language Models to Society (PALMS) with Values-Targeted Datasets, an iterative process to significantly change model behavior by crafting and fine-tuning on a dataset that reflects a predetermined set of target values. We evaluate our pro cess using three metrics: quantitative metrics with human evaluations that score output adherence to a target value, toxicity scoring on outputs; and qualitativ e metrics analyzing the most common word associated with a given social category . Through each iteration, we add additional training dataset examples based on o bserved shortcomings from evaluations. PALMS performs significantly better on al l metrics compared to baseline and control models for a broad range of GPT-3 lan guage model sizes without compromising capability integrity. We find that the ef fectiveness of PALMS increases with model size. We show that significantly adjus ting language model behavior is feasible with a small, hand-curated dataset.
****
The Lazy Online Subgradient Algorithm is Universal on Strongly Convex Domains
Daron Anderson, Douglas Leith
Requests for name changes in the electronic proceedings will be accepted with 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.
****
Computer-Aided Design as Language
Yaroslav Ganin, Sergey Bartunov, Yujia Li, Ethan Keller, Stefano Saliceti

Computer-Aided Design (CAD) applications are used in manufacturing to model ever
ything from coffee mugs to sports cars. These programs are complex and require y
ears of training and experience to master. A component of all CAD models particu
larly difficult to make are the highly structured 2D sketches that lie at the he
art of every 3D construction. In this work, we propose a machine learning model
capable of automatically generating such sketches. Through this, we pave the way
 for developing intelligent tools that would help engineers create better design
s with less effort. The core of our method is a combination of a general-purpose
 language modeling technique alongside an off-the-shelf data serialization proto
col. Additionally, we explore several extensions allowing us to gain finer contr
ol over the generation process. We show that our approach has enough flexibility
 to accommodate the complexity of the domain and performs well for both uncondit
ional synthesis and image-to-sketch translation.
****
COHESIV: Contrastive Object and Hand Embedding Segmentation In Video
Dandan Shan, Richard Higgins, David Fouhey
In this paper we learn to segment hands and hand-held objects from motion. Our s
ystem takes a single RGB image and hand location as input to segment the hand an
d hand-held object. For learning, we generate responsibility maps that show how
well a hand's motion explains other pixels' motion in video. We use these respon
sibility maps as pseudo-labels to train a weakly-supervised neural network using
 an attention-based similarity loss and contrastive loss. Our system outperforms
 alternate methods, achieving good performance on the 100DOH, EPIC-KITCHENS, and
 HO3D datasets.
****
ByPE-VAE: Bayesian Pseudocoresets Exemplar VAE
Qingzhong Ai, LIRONG HE, SHIYU LIU, Zenglin Xu
Recent studies show that advanced priors play a major role in deep generative mo
dels. Exemplar VAE, as a variant of VAE with an exemplar-based prior, has achiev
ed impressive results. However, due to the nature of model design, an exemplar-b
ased model usually requires vast amounts of data to participate in training, whi
ch leads to huge computational complexity. To address this issue, we propose Bay
esian Pseudocoresets Exemplar VAE (ByPE-VAE), a new variant of VAE with a prior
based on Bayesian pseudocoreset. The proposed prior is conditioned on a small-sc
ale pseudocoreset rather than the whole dataset for reducing the computational c
ost and avoiding overfitting. Simultaneously, we obtain the optimal pseudocorese
t via a stochastic optimization algorithm during VAE training aiming to minimize
 the Kullback-Leibler divergence between the prior based on the pseudocoreset an
d that based on the whole dataset. Experimental results show that ByPE-VAE can a
chieve competitive improvements over the state-of-the-art VAEs in the tasks of d
ensity estimation, representation learning, and generative data augmentation. Pa
rticularly, on a basic VAE architecture, ByPE-VAE is up to 3 times faster than E
xemplar VAE while almost holding the performance. Code is available at \url{http
s://github.com/Aiqz/ByPE-VAE}.
****
Recovery Analysis for Plug-and-Play Priors using the Restricted Eigenvalue Condi
tion
Jiaming Liu, Salman Asif, Brendt Wohlberg, Ulugbek Kamilov
The plug-and-play priors (PnP) and regularization by denoising (RED) methods hav
e become widely used for solving inverse problems by leveraging pre-trained deep
 denoisers as image priors. While the empirical imaging performance and the the
oretical convergence properties of these algorithms have been widely investigate
d, their recovery properties have not previously been theoretically analyzed. W
e address this gap by showing how to establish theoretical recovery guarantees f
or PnP/RED by assuming that the solution of these methods lies near the fixed-po
ints of a deep neural network. We also present numerical results comparing the r
ecovery performance of PnP/RED in compressive sensing against that of recent com
pressive sensing algorithms based on generative models. Our numerical results su
ggest that PnP with a pre-trained artifact removal network provides significantl
y better results compared to the existing state-of-the-art methods.

****
Group Equivariant Subsampling
Jin Xu, Hyunjik Kim, Thomas Rainforth, Yee Teh
Subsampling is used in convolutional neural networks (CNNs) in the form of pooling or strided convolutions, to reduce the spatial dimensions of feature maps and to allow the receptive fields to grow exponentially with depth. However, it is known that such subsampling operations are not translation equivariant, unlike convolutions that are translation equivariant. Here, we first introduce translation equivariant subsampling/upsampling layers that can be used to construct exact translation equivariant CNNs. We then generalise these layers beyond translations to general groups, thus proposing group equivariant subsampling/upsampling. We use these layers to construct group equivariant autoencoders (GAEs) that allow us to learn low-dimensional equivariant representations. We empirically verify on images that the representations are indeed equivariant to input translations and rotations, and thus generalise well to unseen positions and orientations. We further use GAEs in models that learn object-centric representations on multi-object datasets, and show improved data efficiency and decomposition compared to non-equivariant baselines.
****
Data Sharing and Compression for Cooperative Networked Control
Jiangnan Cheng, Marco Pavone, Sachin Katti, Sandeep Chinchali, Ao Tang
Sharing forecasts of network timeseries data, such as cellular or electricity load patterns, can improve independent control applications ranging from traffic scheduling to power generation. Typically, forecasts are designed without knowledge of a downstream controller's task objective, and thus simply optimize for mean prediction error. However, such task-agnostic representations are often too large to stream over a communication network and do not emphasize salient temporal features for cooperative control. This paper presents a solution to learn succinct, highly-compressed forecasts that are co-designed with a modular controller's task objective. Our simulations with real cellular, Internet-of-Things (IoT), and electricity load data show we can improve a model predictive controller's performance by at least 25% while transmitting 80% less data than the competing method. Further, we present theoretical compression results for a networked variant of the classical linear quadratic regulator (LQR) control problem.
****
Hyperbolic Procrustes Analysis Using Riemannian Geometry
Ya-Wei Eileen Lin, Yuval Kluger, Ronen Talmon
Label-free alignment between datasets collected at different times, locations, or by different instruments is a fundamental scientific task. Hyperbolic spaces have recently provided a fruitful foundation for the development of informative representations of hierarchical data. Here, we take a purely geometric approach for label-free alignment of hierarchical datasets and introduce hyperbolic Procrustes analysis (HPA). HPA consists of new implementations of the three prototypical Procrustes analysis components: translation, scaling, and rotation, based on the Riemannian geometry of the Lorentz model of hyperbolic space. We analyze the proposed components, highlighting their useful properties for alignment. The efficacy of HPA, its theoretical properties, stability and computational efficiency are demonstrated in simulations. In addition, we showcase its performance on three batch correction tasks involving gene expression and mass cytometry data. Specifically, we demonstrate high-quality unsupervised batch effect removal from data acquired at different sites and with different technologies that outperforms recent methods for label-free alignment in hyperbolic spaces.
****
No Fear of Heterogeneity: Classifier Calibration for Federated Learning with Non-IID Data
Mi Luo, Fei Chen, Dapeng Hu, Yifan Zhang, Jian Liang, Jiashi Feng
A central challenge in training classification models in the real-world federated system is learning with non-IID data. To cope with this, most of the existing works involve enforcing regularization in local optimization or improving the model aggregation scheme at the server. Other works also share public datasets or

synthesized samples to supplement the training of under-represented classes or i
ntroduce a certain level of personalization. Though effective, they lack a deep
understanding of how the data heterogeneity affects each layer of a deep classif
ication model. In this paper, we bridge this gap by performing an experimental a
nalysis of the representations learned by different layers. Our observations are
 surprising: (1) there exists a greater bias in the classifier than other layers
, and (2) the classification performance can be significantly improved by post-c
alibrating the classifier after federated training. Motivated by the above findi
ngs, we propose a novel and simple algorithm called Classifier Calibration with
Virtual Representations (CCVR), which adjusts the classifier using virtual repre
sentations sampled from an approximated gaussian mixture model. Experimental res
ults demonstrate that CCVR achieves state-of-the-art performance on popular fede
rated learning benchmarks including CIFAR-10, CIFAR-100, and CINIC-10. We hope t
hat our simple yet effective method can shed some light on the future research o
f federated learning with non-IID data.
****
Preconditioned Gradient Descent for Over-Parameterized Nonconvex Matrix Factoriz
ation
Jialun Zhang, Salar Fattahi, Richard Y 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.
****
Improving Contrastive Learning on Imbalanced Data via Open-World Sampling
Ziyu Jiang, Tianlong Chen, Ting Chen, Zhangyang Wang
Contrastive learning approaches have achieved great success in learning visual r
epresentations with few labels of the target classes. That implies a tantalizing
 possibility of scaling them up beyond a curated “seed" benchmark, to incorporat
ing more unlabeled images from the internet-scale external sources to enhance it
s performance. However, in practice, larger amount of unlabeled data will requir
e more computing resources due to the bigger model size and longer training need
ed. Moreover, open-world unlabeled data usually follows an implicit long-tail cl
ass or attribute distribution, many of which also do not belong to the target cl
asses. Blindly leveraging all unlabeled data hence can lead to the data imbalanc
e as well as distraction issues. This motivates us to seek a principled approach
 to strategically select unlabeled data from an external source, in order to lea
rn generalizable, balanced and diverse representations for relevant classes. In
this work, we present an open-world unlabeled data sampling framework called Mod
el-Aware K-center (MAK), which follows three simple principles: (1) tailness, wh
ich encourages sampling of examples from tail classes, by sorting the empirical
contrastive loss expectation (ECLE) of samples over random data augmentations; (
2) proximity, which rejects the out-of-distribution outliers that may distract t
raining; and (3) diversity, which ensures diversity in the set of sampled exampl
es. Empirically, using ImageNet-100-LT (without labels) as the seed dataset and
two “noisy” external data sources, we demonstrate that MAK can consistently impr
ove both the overall representation quality and the class balancedness of the le
arned features, as evaluated via linear classifier evaluation on full-shot and f
ew-shot settings. Thecode is available at: https://github.com/VITA-Group/MAK.
****
Searching for Efficient Transformers for Language Modeling
David So, Wojciech Ma■ke, Hanxiao Liu, Zihang Dai, Noam Shazeer, Quoc V Le
Large Transformer models have been central to recent advances in natural languag
e processing. The training and inference costs of these models, however, have gr
own rapidly and become prohibitively expensive. Here we aim to reduce the costs
of Transformers by searching for a more efficient variant. Compared to previous
approaches, our search is performed at a lower level, over the primitives that d
efine a Transformer TensorFlow program. We identify an architecture, named Prime
r, that has a smaller training cost than the original Transformer and other vari
ants for auto-regressive language modeling. Primer’s improvements can be mostly

attributed to two simple modifications: squaring ReLU activations and adding a depthwise convolution layer after each Q, K, and V projection in self-attention.Experiments show Primer's gains over Transformer increase as compute scale grows and follow a power law with respect to quality at optimal model sizes. We also verify empirically that Primer can be dropped into different codebases to significantly speed up training without additional tuning. For example, at a 500M parameter size, Primer improves the original T5 architecture on C4 auto-regressive language modeling, reducing the training cost by 4X. Furthermore, the reduced training cost means Primer needs much less compute to reach a target one-shot performance. For instance, in a 1.9B parameter configuration similar to GPT-3 XL, Primer uses 1/3 of the training compute to achieve the same one-shot performance as Transformer. We open source our models and several comparisons in T5 to help with reproducibility.
****
Scaling Ensemble Distribution Distillation to Many Classes with Proxy Targets
Max Ryabinin, Andrey Malinin, Mark Gales
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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-Person 3D Motion Prediction with Multi-Range Transformers
Jiashun Wang, Huazhe Xu, Medhini Narasimhan, Xiaolong Wang
We propose a novel framework for multi-person 3D motion trajectory prediction. Our key observation is that a human's action and behaviors may highly depend on the other persons around. Thus, instead of predicting each human pose trajectory in isolation, we introduce a Multi-Range Transformers model which contains of a local-range encoder for individual motion and a global-range encoder for social interactions. The Transformer decoder then performs prediction for each person by taking a corresponding pose as a query which attends to both local and global-range encoder features. Our model not only outperforms state-of-the-art methods on long-term 3D motion prediction, but also generates diverse social interactions. More interestingly, our model can even predict 15-person motion simultaneously by automatically dividing the persons into different interaction groups. Project page with code is available at https://jiashunwang.github.io/MRT/.
****
STEM: A Stochastic Two-Sided Momentum Algorithm Achieving Near-Optimal Sample and Communication Complexities for Federated Learning
Prashant Khanduri, PRANAY SHARMA, Haibo Yang, Mingyi Hong, Jia Liu, Ketan Rajawat, Pramod Varshney
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Bubblewrap: Online tiling and real-time flow prediction on neural manifolds
Anne Draelos, Pranjal Gupta, Na Young Jun, Chaichontat Sriworarat, John Pearson
While most classic studies of function in experimental neuroscience have focused on the coding properties of individual neurons, recent developments in recording technologies have resulted in an increasing emphasis on the dynamics of neural populations. This has given rise to a wide variety of models for analyzing population activity in relation to experimental variables, but direct testing of many neural population hypotheses requires intervening in the system based on current neural state, necessitating models capable of inferring neural state online. Existing approaches, primarily based on dynamical systems, require strong parametric assumptions that are easily violated in the noise-dominated regime and do not scale well to the thousands of data channels in modern experiments. To address this problem, we propose a method that combines fast, stable dimensionality reduction with a soft tiling of the resulting neural manifold, allowing dynamics to be approximated as a probability flow between tiles. This method can be fit ef

ficiently using online expectation maximization, scales to tens of thousands of tiles, and outperforms existing methods when dynamics are noise-dominated or feature multi-modal transition probabilities. The resulting model can be trained at kiloHertz data rates, produces accurate approximations of neural dynamics within minutes, and generates predictions on submillisecond time scales. It retains predictive performance throughout many time steps into the future and is fast enough to serve as a component of closed-loop causal experiments.
****
The Semi-Random Satisfaction of Voting Axioms
Lirong Xia
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Deep Marching Tetrahedra: a Hybrid Representation for High-Resolution 3D Shape Synthesis
Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, Sanja Fidler
We introduce DMTet, a deep 3D conditional generative model that can synthesize high-resolution 3D shapes using simple user guides such as coarse voxels. It marries the merits of implicit and explicit 3D representations by leveraging a novel hybrid 3D representation. Compared to the current implicit approaches, which are trained to regress the signed distance values, DMTet directly optimizes for the reconstructed surface, which enables us to synthesize finer geometric details with fewer artifacts. Unlike deep 3D generative models that directly generate explicit representations such as meshes, our model can synthesize shapes with arbitrary topology. The core of DMTet includes a deformable tetrahedral grid that encodes a discretized signed distance function and a differentiable marching tetrahedra layer that converts the implicit signed distance representation to the explicit surface mesh representation. This combination allows joint optimization of the surface geometry and topology as well as generation of the hierarchy of subdivisions using reconstruction and adversarial losses defined explicitly on the surface mesh. Our approach significantly outperforms existing work on conditional shape synthesis from coarse voxel inputs, trained on a dataset of complex 3D animal shapes. Project page: https://nv-tlabs.github.io/DMTet/.
****
Learning to Combine Per-Example Solutions for Neural Program Synthesis
Disha Shrivastava, Hugo Larochelle, Daniel Tarlow
The goal of program synthesis from examples is to find a computer program that is consistent with a given set of input-output examples. Most learning-based approaches try to find a program that satisfies all examples at once. Our work, by contrast, considers an approach that breaks the problem into two stages: (a) find programs that satisfy only one example, and (b) leverage these per-example solutions to yield a program that satisfies all examples. We introduce the Cross Aggregator neural network module based on a multi-head attention mechanism that learns to combine the cues present in these per-example solutions to synthesize a global solution. Evaluation across programs of different lengths and under two different experimental settings reveal that when given the same time budget, our technique significantly improves the success rate over PCCoder [Zohar et. al 2018] and other ablation baselines.
****
On Success and Simplicity: A Second Look at Transferable Targeted Attacks
Zhengyu Zhao, Zhuoran Liu, Martha Larson
Achieving transferability of targeted attacks is reputed to be remarkably difficult. The current state of the art has resorted to resource-intensive solutions that necessitate training model(s) for each target class with additional data. In our investigation, we find, however, that simple transferable attacks which require neither model training nor additional data can achieve surprisingly strong targeted transferability. This insight has been overlooked until now, mainly because the widespread practice of attacking with only few iterations has largely l

imited the attack convergence to optimal targeted transferability. In particular
, we, for the first time, identify that a very simple logit loss can largely sur
pass the commonly adopted cross-entropy loss, and yield even better results than
 the resource-intensive state of the art. Our analysis spans a variety of transf
er scenarios, especially including three new, realistic scenarios: an ensemble t
ransfer scenario with little model similarity, a worse-case scenario with low-ra
nked target classes, and also a real-world attack on the Google Cloud Vision API
. Results in these new transfer scenarios demonstrate that the commonly adopted,
 easy scenarios cannot fully reveal the actual strength of different attacks and
 may cause misleading comparative results. We also show the usefulness of the si
mple logit loss for generating targeted universal adversarial perturbations in a
 data-free manner. Overall, the aim of our analysis is to inspire a more meaning
ful evaluation on targeted transferability. Code is available at https://github.
com/ZhengyuZhao/Targeted-Tansfer.
****
Provably efficient, succinct, and precise explanations
Guy Blanc, 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.
****
Refined Learning Bounds for Kernel and Approximate k-Means
Yong 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.
****
Learning Causal Semantic Representation for Out-of-Distribution Prediction
Chang Liu, Xinwei Sun, Jindong Wang, Haoyue Tang, Tao Li, Tao Qin, Wei Chen, Tie
-Yan Liu
Conventional supervised learning methods, especially deep ones, are found to be
sensitive to out-of-distribution (OOD) examples, largely because the learned rep
resentation mixes the semantic factor with the variation factor due to their dom
ain-specific correlation, while only the semantic factor causes the output. To a
ddress the problem, we propose a Causal Semantic Generative model (CSG) based on
 a causal reasoning so that the two factors are modeled separately, and develop
methods for OOD prediction from a single training domain, which is common and ch
allenging. The methods are based on the causal invariance principle, with a nove
l design in variational Bayes for both efficient learning and easy prediction. T
heoretically, we prove that under certain conditions, CSG can identify the seman
tic factor by fitting training data, and this semantic-identification guarantees
 the boundedness of OOD generalization error and the success of adaptation. Empi
rical study shows improved OOD performance over prevailing baselines.
****
A first-order primal-dual method with adaptivity to local smoothness
Maria-Luiza Vladarean, Yura Malitsky, 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.
****
A Theory-Driven Self-Labeling Refinement Method for Contrastive Representation L
earning
Pan Zhou, Caiming Xiong, Xiaotong Yuan, Steven Chu Hong Hoi
For an image query, unsupervised contrastive learning labels crops of the sam
e image as positives, and other image crops as negatives. Although intuitive,
such a native label assignment strategy cannot reveal the underlying semantic si
milarity between a query and its positives and negatives, and impairs performa

nce, since some negatives are semantically similar to the query or even share the same semantic class as the query. In this work, we first prove that for contrastive learning, inaccurate label assignment heavily impairs its generalization for semantic instance discrimination, while accurate labels benefit its generalization. Inspired by this theory, we propose a novel self-labeling refinement approach for contrastive learning. It improves the label quality via two complementary modules: (i) self-labeling refinery (SLR) to generate accurate labels and (ii) momentum mixup (MM) to enhance similarity between query and its positive. SLR uses a positive of a query to estimate semantic similarity between a query and its positive and negatives, and combines estimated similarity with vanilla label assignment in contrastive learning to iteratively generate more accurate and informative soft labels. We theoretically show that our SLR can exactly recover the true semantic labels of label-corrupted data, and supervises networks to achieve zero prediction error on classification tasks. MM randomly combines queries and positives to increase semantic similarity between the generated virtual queries and their positives so as to improves label accuracy. Experimental results on CIFAR10, ImageNet, VOC and COCO show the effectiveness of our method.
****
Adversarial Robustness with Semi-Infinite Constrained Learning
Alexander Robey, Luiz Chamon, George J. Pappas, Hamed Hassani, Alejandro Ribeiro
Despite strong performance in numerous applications, the fragility of deep learning to input perturbations has raised serious questions about its use in safety-critical domains. While adversarial training can mitigate this issue in practice, state-of-the-art methods are increasingly application-dependent, heuristic in nature, and suffer from fundamental trade-offs between nominal performance and robustness. Moreover, the problem of finding worst-case perturbations is non-convex and underparameterized, both of which engender a non-favorable optimization landscape. Thus, there is a gap between the theory and practice of robust learning, particularly with respect to when and why adversarial training works. In this paper, we take a constrained learning approach to address these questions and to provide a theoretical foundation for robust learning. In particular, we leverage semi-infinite optimization and non-convex duality theory to show that adversarial training is equivalent to a statistical problem over perturbation distributions. Notably, we show that a myriad of previous robust training techniques can be recovered for particular, sub-optimal choices of these distributions. Using these insights, we then propose a hybrid Langevin Markov Chain Monte Carlo approach for which several common algorithms (e.g., PGD) are special cases. Finally, we show that our approach can mitigate the trade-off between nominal and robust performance, yielding state-of-the-art results on MNIST and CIFAR-10. Our code is available at: https://github.com/arobey1/advbench.
****
Conformal Time-series Forecasting
Kamile Stankeviciute, Ahmed M. Alaa, Mihaela van der Schaar
Current approaches for multi-horizon time series forecasting using recurrent neural networks (RNNs) focus on issuing point estimates, which is insufficient for decision-making in critical application domains where an uncertainty estimate is also required. Existing approaches for uncertainty quantification in RNN-based time-series forecasts are limited as they may require significant alterations to the underlying model architecture, may be computationally complex, may be difficult to calibrate, may incur high sample complexity, and may not provide theoretical guarantees on frequentist coverage. In this paper, we extend the inductive conformal prediction framework to the time-series forecasting setup, and propose a lightweight algorithm to address all of the above limitations, providing uncertainty estimates with theoretical guarantees for any multi-horizon forecast predictor and any dataset with minimal exchangeability assumptions. We demonstrate the effectiveness of our approach by comparing it with existing benchmarks on a variety of synthetic and real-world datasets.
****
A 3D Generative Model for Structure-Based Drug Design

Shitong Luo, Jiaqi Guan, Jianzhu Ma, Jian Peng
We study a fundamental problem in structure-based drug design --- generating molecules that bind to specific protein binding sites. While we have witnessed the great success of deep generative models in drug design, the existing methods are mostly string-based or graph-based. They are limited by the lack of spatial information and thus unable to be applied to structure-based design tasks. Particularly, such models have no or little knowledge of how molecules interact with their target proteins exactly in 3D space. In this paper, we propose a 3D generative model that generates molecules given a designated 3D protein binding site. Specifically, given a binding site as the 3D context, our model estimates the probability density of atom's occurrences in 3D space --- positions that are more likely to have atoms will be assigned higher probability. To generate 3D molecules, we propose an auto-regressive sampling scheme --- atoms are sampled sequentially from the learned distribution until there is no room for new atoms. Combined with this sampling scheme, our model can generate valid and diverse molecules, which could be applicable to various structure-based molecular design tasks such as molecule sampling and linker design. Experimental results demonstrate that molecules sampled from our model exhibit high binding affinity to specific targets and good drug properties such as drug-likeness even if the model is not explicitly optimized for them.
****
Bootstrapping the Error of Oja's Algorithm
Robert Lunde, Purnamrita Sarkar, Rachel Ward
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Landscape analysis of an improved power method for tensor decomposition
Joe Kileel, Timo Klock, João M Pereira
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 Offline Imitating Learning
Minghuan Liu, Hanye Zhao, Zhengyu Yang, Jian Shen, Weinan Zhang, Li Zhao, Tie-Yan Liu
Offline reinforcement learning (RL) tasks require the agent to learn from a pre-collected dataset with no further interactions with the environment. Despite the potential to surpass the behavioral policies, RL-based methods are generally impractical due to the training instability and bootstrapping the extrapolation errors, which always require careful hyperparameter tuning via online evaluation. In contrast, offline imitation learning (IL) has no such issues since it learns the policy directly without estimating the value function by bootstrapping. However, IL is usually limited in the capability of the behavioral policy and tends to learn a mediocre behavior from the dataset collected by the mixture of policies. In this paper, we aim to take advantage of IL but mitigate such a drawback. Observing that behavior cloning is able to imitate neighboring policies with less data, we propose \textit{Curriculum Offline Imitation Learning (COIL)}, which utilizes an experience picking strategy to make the agent imitate from adaptive neighboring policies with a higher return, and improves the current policy along curriculum stages. On continuous control benchmarks, we compare COIL against both imitation-based methods and RL-based methods, showing that COIL not only avoids just learning a mediocre behavior on mixed datasets but is also even competitive with state-of-the-art offline RL methods.
****
Robust Pose Estimation in Crowded Scenes with Direct Pose-Level Inference
Dongkai Wang, Shiliang Zhang, Gang Hua
Multi-person pose estimation in crowded scenes is challenging because overlappin

g and occlusions make it difficult to detect person bounding boxes and infer pose cues from individual keypoints. To address those issues, this paper proposes a direct pose-level inference strategy that is free of bounding box detection and keypoint grouping. Instead of inferring individual keypoints, the Pose-level Inference Network (PINet) directly infers the complete pose cues for a person from his/her visible body parts. PINet first applies the Part-based Pose Generation (PPG) to infer multiple coarse poses for each person from his/her body parts. Those coarse poses are refined by the Pose Refinement module through incorporating pose priors, and finally are fused in the Pose Fusion module. PINet relies on discriminative body parts to differentiate overlapped persons, and applies visual body cues to infer the global pose cues. Experiments on several crowded scenes pose estimation benchmarks demonstrate the superiority of PINet. For instance, it achieves 59.8% AP on the OCHuman dataset, outperforming the recent works by a large margin.
****
Ising Model Selection Using ℓ_{1}-Regularized Linear Regression: A Statistical Mechanics Analysis
Xiangming Meng, Tomoyuki Obuchi, Yoshiyuki Kabashima
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 Prediction using Conditional Histograms
Matteo Sesia, Yaniv Romano
This paper develops a conformal method to compute prediction intervals for non-parametric regression that can automatically adapt to skewed data. Leveraging black-box machine learning algorithms to estimate the conditional distribution of the outcome using histograms, it translates their output into the shortest prediction intervals with approximate conditional coverage. The resulting prediction intervals provably have marginal coverage in finite samples, while asymptotically achieving conditional coverage and optimal length if the black-box model is consistent. Numerical experiments with simulated and real data demonstrate improved performance compared to state-of-the-art alternatives, including conformalized quantile regression and other distributional conformal prediction approaches.
****
Contrastive Graph Poisson Networks: Semi-Supervised Learning with Extremely Limited Labels
Sheng Wan, Yibing Zhan, Liu Liu, Baosheng Yu, Shirui Pan, Chen Gong
Graph Neural Networks (GNNs) have achieved remarkable performance in the task of semi-supervised node classification. However, most existing GNN models require sufficient labeled data for effective network training. Their performance can be seriously degraded when labels are extremely limited. To address this issue, we propose a new framework termed Contrastive Graph Poisson Networks (CGPN) for node classification under extremely limited labeled data. Specifically, our CGPN derives from variational inference; integrates a newly designed Graph Poisson Network (GPN) to effectively propagate the limited labels to the entire graph and a normal GNN, such as Graph Attention Network, that flexibly guides the propagation of GPN; applies a contrastive objective to further exploit the supervision information from the learning process of GPN and GNN models. Essentially, our CGPN can enhance the learning performance of GNNs under extremely limited labels by contrastively propagating the limited labels to the entire graph. We conducted extensive experiments on different types of datasets to demonstrate the superiority of CGPN.
****
Collaborative Uncertainty in Multi-Agent Trajectory Forecasting
Bohan Tang, Yiqi Zhong, Ulrich Neumann, Gang Wang, Siheng Chen, Ya Zhang
Uncertainty modeling is critical in trajectory-forecasting systems for both interpretation and safety reasons. To better predict the future trajectories of multiple agents, recent works have introduced interaction modules to capture interac

tions among agents. This approach leads to correlations among the predicted traj
ectories. However, the uncertainty brought by such correlations is neglected. To
 fill this gap, we propose a novel concept, collaborative uncertainty (CU), whic
h models the uncertainty resulting from the interaction module. We build a gener
al CU-based framework to make a prediction model learn the future trajectory and
 the corresponding uncertainty. The CU-based framework is integrated as a plugin
 module to current state-of-the-art (SOTA) systems and deployed in two special c
ases based on multivariate Gaussian and Laplace distributions. In each case, we
conduct extensive experiments on two synthetic datasets and two public, large-sc
ale benchmarks of trajectory forecasting. The results are promising: 1) The resu
lts of synthetic datasets show that CU-based framework allows the model to nicel
y rebuild the ground-truth distribution. 2) The results of trajectory forecastin
g benchmarks demonstrate that the CU-based framework steadily helps SOTA systems
 improve their performances. Specially, the proposed CU-based framework helps Ve
ctorNet improve by 57 cm regarding Final Displacement Error on nuScenes dataset.
 3) The visualization results of CU illustrate that the value of CU is highly re
lated to the amount of the interactive information among agents.
****
Network-to-Network Regularization: Enforcing Occam's Razor to Improve Generaliza
tion
Rohan Ghosh, Mehul Motani
What makes a classifier have the ability to generalize? There have been a lot of
 important attempts to address this question, but a clear answer is still elusiv
e. Proponents of complexity theory find that the complexity of the classifier's
function space is key to deciding generalization, whereas other recent work reve
als that classifiers which extract invariant feature representations are likely
to generalize better. Recent theoretical and empirical studies, however, have sh
own that even within a classifier's function space, there can be significant dif
ferences in the ability to generalize. Specifically, empirical studies have show
n that among functions which have a good training data fit, functions with lower
 Kolmogorov complexity (KC) are likely to generalize better, while the opposite
is true for functions of higher KC. Motivated by these findings, we propose, in
this work, a novel measure of complexity called Kolmogorov Growth (KG), which we
 use to derive new generalization error bounds that only depend on the final cho
ice of the classification function. Guided by the bounds, we propose a novel way
 of regularizing neural networks by constraining the network trajectory to remai
n in the low KG zone during training. Minimizing KG while learning is akin to ap
plying the Occam's razor to neural networks. The proposed approach, called netwo
rk-to-network regularization, leads to clear improvements in the generalization
ability of classifiers. We verify this for three popular image datasets (MNIST,
CIFAR-10, CIFAR-100) across varying training data sizes. Empirical studies find
that conventional training of neural networks, unlike network-to-network regular
ization, leads to networks of high KG and lower test accuracies. Furthermore, we
 present the benefits of N2N regularization in the scenario where the training d
ata labels are noisy. Using N2N regularization, we achieve competitive performan
ce on MNIST, CIFAR-10 and CIFAR-100 datasets with corrupted training labels, sig
nificantly improving network performance compared to standard cross-entropy base
lines in most cases. These findings illustrate the many benefits obtained from i
mposing a function complexity prior like Kolmogorov Growth during the training p
rocess.
****
Generalized and Discriminative Few-Shot Object Detection via SVD-Dictionary Enha
ncement
Aming WU, Suqi Zhao, Cheng Deng, Wei Liu
Few-shot object detection (FSOD) aims to detect new objects based on few annotat
ed samples. To alleviate the impact of few samples, enhancing the generalization
 and discrimination abilities of detectors on new objects plays an important rol
e. In this paper, we explore employing Singular Value Decomposition (SVD) to boo
st both the generalization and discrimination abilities. In specific, we propose
 a novel method, namely, SVD-Dictionary enhancement, to build two separated spac

es based on the sorted singular values. Concretely, the eigenvectors correspondi ng to larger singular values are used to build the generalization space in which localization is performed, as these eigenvectors generally suppress certain var iations (e.g., the variation of styles) and contain intrinsical characteristics of objects. Meanwhile, since the eigenvectors corresponding to relatively smalle r singular values may contain richer category-related information, we can utiliz e them to build the discrimination space in which classification is performed. D ictionary learning is further leveraged to capture high-level discriminative inf ormation from the discrimination space, which is beneficial for improving detect ion accuracy. In the experiments, we separately verify the effectiveness of our method on PASCAL VOC and COCO benchmarks. Particularly, for the 2-shot case in V OC split1, our method significantly outperforms the baseline by 6.2\%. Moreover, visualization analysis shows that our method is instrumental in doing FSOD.
****
Conditioning Sparse Variational Gaussian Processes for Online Decision-making
Wesley J. Maddox, Samuel Stanton, 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-auth ors prior to requesting a name change in the electronic proceedings.
****
Spherical Motion Dynamics: Learning Dynamics of Normalized Neural Network using SGD and Weight Decay
Ruosi Wan, Zhanxing Zhu, Xiangyu Zhang, Jian Sun
In this paper, we comprehensively reveal the learning dynamics of normalized neu ral network using Stochastic Gradient Descent (with momentum) and Weight Decay (WD), named as Spherical Motion Dynamics (SMD). Most related works focus on study ing behavior of effective learning rate" inequilibrium" state, i.e. assuming wei ght norm remains unchanged. However, their discussion on why this equilibrium ca n be reached is either absent or less convincing. Our work directly explores the cause of equilibrium, as a special state of SMD. Specifically, 1) we introduce the assumptions that can lead to equilibrium state in SMD, and prove equilibrium can be reached in a linear rate regime under given assumptions; 2) we propose ` `angular update" as a substitute for effective learning rate to depict the state of SMD, and derive the theoretical value of angular update in equilibrium state ; 3) we verify our assumptions and theoretical results on various large-scale co mputer vision tasks including ImageNet and MSCOCO with standard settings. Experi ment results show our theoretical findings agree well with empirical observation s. We also show that the behavior of angular update in SMD can produce interesti ng effect to the optimization of neural network in practice.
****
Imitating Deep Learning Dynamics via Locally Elastic Stochastic Differential Eq uations
Jiayao Zhang, Hua Wang, Weijie Su
Understanding the training dynamics of deep learning models is perhaps a necessa ry step toward demystifying the effectiveness of these models. In particular, ho w do training data from different classes gradually become separable in their fe ature spaces when training neural networks using stochastic gradient descent? In this paper, we model the evolution of features during deep learning training us ing a set of stochastic differential equations (SDEs) that each corresponding to a training sample. As a crucial ingredient in our modeling strategy, each SDE c ontains a drift term that reflects the impact of backpropagation at an input on the features of all samples. Our main finding uncovers a sharp phase transition phenomenon regarding the intra-class impact: if the SDEs are locally elastic in the sense that the impact is more significant on samples from the same class as the input, the features of training data become linearly separable---meaning van ishing training loss; otherwise, the features are not separable, no matter how l ong the training time is. In the presence of local elasticity, moreover, an anal ysis of our SDEs shows the emergence of a simple geometric structure called neur al collapse of the features. Taken together, our results shed light on the decis

ive role of local elasticity underlying the training dynamics of neural networks. We corroborate our theoretical analysis with experiments on a synthesized dataset of geometric shapes as well as on CIFAR-10.
****

Probabilistic Forecasting: A Level-Set Approach
Hilaf Hasson, Bernie Wang, Tim Januschowski, Jan Gasthaus
Large-scale time series panels have become ubiquitous over the last years in areas such as retail, operational metrics, IoT, and medical domain (to name only a few). This has resulted in a need for forecasting techniques that effectively leverage all available data by learning across all time series in each panel. Among the desirable properties of forecasting techniques, being able to generate probabilistic predictions ranks among the top. In this paper, we therefore present Level Set Forecaster (LSF), a simple yet effective general approach to transform a point estimator into a probabilistic one. By recognizing the connection of our algorithm to random forests (RFs) and quantile regression forests (QRFs), we are able to prove consistency guarantees of our approach under mild assumptions on the underlying point estimator. As a byproduct, we prove the first consistency results for QRFs under the CART-splitting criterion. Empirical experiments show that our approach, equipped with tree-based models as the point estimator, rivals state-of-the-art deep learning models in terms of forecasting accuracy.
****

Roto-translated Local Coordinate Frames For Interacting Dynamical Systems
Miltiadis Kofinas, Naveen Nagaraja, Efstratios Gavves
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

ParK: Sound and Efficient Kernel Ridge Regression by Feature Space Partitions
Luigi Carratino, Stefano Vigogna, Daniele Calandriello, Lorenzo Rosasco
We introduce ParK, a new large-scale solver for kernel ridge regression. Our approach combines partitioning with random projections and iterative optimization to reduce space and time complexity while provably maintaining the same statistical accuracy. In particular, constructing suitable partitions directly in the feature space rather than in the input space, we promote orthogonality between the local estimators, thus ensuring that key quantities such as local effective dimension and bias remain under control. We characterize the statistical-computational tradeoff of our model, and demonstrate the effectiveness of our method by numerical experiments on large-scale datasets.
****

Scaling Gaussian Processes with Derivative Information Using Variational Inference
Misha Padidar, Xinran Zhu, Leo Huang, Jacob Gardner, David Bindel
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 Representation of Solutions to Elliptic PDEs in Barron Spaces
Ziang Chen, Jianfeng Lu, Yulong 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.
****

A/B Testing for Recommender Systems in a Two-sided Marketplace
Preetam Nandy, Divya Venugopalan, Chun Lo, Shaunak Chatterjee
Two-sided marketplaces are standard business models of many online platforms (e.g., Amazon, Facebook, LinkedIn), wherein the platforms have consumers, buyers or content viewers on one side and producers, sellers or content-creators on the o

ther. Consumer side measurement of the impact of a treatment variant can be done via simple online A/B testing. Producer side measurement is more challenging because the producer experience depends on the treatment assignment of the consumers. Existing approaches for producer side measurement are either based on graph cluster-based randomization or on certain treatment propagation assumptions. The former approach results in low-powered experiments as the producer-consumer network density increases and the latter approach lacks a strict notion of error control. In this paper, we propose (i) a quantification of the quality of a producer side experiment design, and (ii) a new experiment design mechanism that generates high-quality experiments based on this quantification. Our approach, called UniCoRn (Unifying Counterfactual Rankings), provides explicit control over the quality of the experiment and its computation cost. Further, we prove that our experiment design is optimal to the proposed design quality measure. Our approach is agnostic to the density of the producer-consumer network and does not rely on any treatment propagation assumption. Moreover, unlike the existing approaches, we do not need to know the underlying network in advance, making this widely applicable to the industrial setting where the underlying network is unknown and challenging to predict a priori due to its dynamic nature. We use simulations to validate our approach and compare it against existing methods. We also deployed UniCoRn in an edge recommendation application that serves tens of millions of members and billions of edge recommendations daily.
****
Retiring Adult: New Datasets for Fair Machine Learning
Frances Ding, Moritz Hardt, John Miller, Ludwig Schmidt
Although the fairness community has recognized the importance of data, researchers in the area primarily rely on UCI Adult when it comes to tabular data. Derived from a 1994 US Census survey, this dataset has appeared in hundreds of research papers where it served as the basis for the development and comparison of many algorithmic fairness interventions. We reconstruct a superset of the UCI Adult data from available US Census sources and reveal idiosyncrasies of the UCI Adult dataset that limit its external validity. Our primary contribution is a suite of new datasets derived from US Census surveys that extend the existing data ecosystem for research on fair machine learning. We create prediction tasks relating to income, employment, health, transportation, and housing. The data span multiple years and all states of the United States, allowing researchers to study temporal shift and geographic variation. We highlight a broad initial sweep of new empirical insights relating to trade-offs between fairness criteria, performance of algorithmic interventions, and the role of distribution shift based on our new datasets. Our findings inform ongoing debates, challenge some existing narratives, and point to future research directions.
****
Cardinality constrained submodular maximization for random streams
Paul Liu, Aviad Rubinstein, Jan Vondrak, Junyao Zhao
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Self-Instantiated Recurrent Units with Dynamic Soft Recursion
Aston Zhang, Yi Tay, Yikang Shen, Alvin Chan, SHUAI ZHANG
While standard recurrent neural networks explicitly impose a chain structure on different forms of data, they do not have an explicit bias towards recursive self-instantiation where the extent of recursion is dynamic. Given diverse and even growing data modalities (e.g., logic, algorithmic input and output, music, code, images, and language) that can be expressed in sequences and may benefit from more architectural flexibility, we propose the self-instantiated recurrent unit (Self-IRU) with a novel inductive bias towards dynamic soft recursion. On one hand, theSelf-IRU is characterized by recursive self-instantiation via its gating functions, i.e., gating mechanisms of the Self-IRU are controlled by instances of the Self-IRU itself, which are repeatedly invoked in a recursive fashion. On

the other hand, the extent of the Self-IRU recursion is controlled by gates whos
e values are between 0 and 1 and may vary across the temporal dimension of seque
nces, enabling dynamic soft recursion depth at each time step. The architectura
l flexibility and effectiveness of our proposed approach are demonstrated across
 multiple data modalities. For example, the Self-IRU achieves state-of-the-art p
erformance on the logical inference dataset [Bowman et al., 2014] even when comp
aring with competitive models that have access to ground-truth syntactic informa
tion.
****
Sparse Uncertainty Representation in Deep Learning with Inducing Weights
Hippolyt Ritter, Martin Kukla, Cheng Zhang, Yingzhen 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.
****
Scalable Inference of Sparsely-changing Gaussian Markov Random Fields
Salar Fattahi, Andres Gomez
Requests for name changes in the electronic proceedings will be accepted with 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.
****
Grad2Task: Improved Few-shot Text Classification Using Gradients for Task Repres
entation
Jixuan Wang, Kuan-Chieh Wang, Frank Rudzicz, Michael Brudno
Large pretrained language models (LMs) like BERT have improved performance in ma
ny disparate natural language processing (NLP) tasks. However, fine tuning such
models requires a large number of training examples for each target task. Simult
aneously, many realistic NLP problems are "few shot", without a sufficiently lar
ge training set. In this work, we propose a novel conditional neural process-bas
ed approach for few-shot text classification that learns to transfer from other
diverse tasks with rich annotation. Our key idea is to represent each task using
 gradient information from a base model and to train an adaptation network that
modulates a text classifier conditioned on the task representation. While previo
us task-aware few-shot learners represent tasks by input encoding, our novel tas
k representation is more powerful, as the gradient captures input-output relatio
nships of a task. Experimental results show that our approach outperforms tradit
ional fine-tuning, sequential transfer learning, and state-of-the-art meta learn
ing approaches on a collection of diverse few-shot tasks. We further conducted a
nalysis and ablations to justify our design choices.
****
Learnability of Linear Thresholds from Label Proportions
Rishi Saket
Requests for name changes in the electronic proceedings will be accepted with 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 variational approximate posterior for the deep Wishart process
Sebastian Ober, Laurence Aitchison
Recent work introduced deep kernel processes as an entirely kernel-based alterna
tive to NNs (Aitchison et al. 2020). Deep kernel processes flexibly learn good t
op-layer representations by alternately sampling the kernel from a distribution
over positive semi-definite matrices and performing nonlinear transformations. A
 particular deep kernel process, the deep Wishart process (DWP), is of particula
r interest because its prior can be made equivalent to deep Gaussian process (DG
P) priors for kernels that can be expressed entirely in terms of Gram matrices.
However, inference in DWPs has not yet been possible due to the lack of sufficie
ntly flexible distributions over positive semi-definite matrices. Here, we give

a novel approach to obtaining flexible distributions over positive semi-definite
 matrices by generalising the Bartlett decomposition of the Wishart probability
density. We use this new distribution to develop an approximate posterior for th
e DWP that includes dependency across layers. We develop a doubly-stochastic ind
ucing-point inference scheme for the DWP and show experimentally that inference
in the DWP can improve performance over doing inference in a DGP with the equiva
lent prior.
****
Neural Pseudo-Label Optimism for the Bank Loan Problem
Aldo Pacchiano, Shaun Singh, Edward Chou, Alex Berg, Jakob Foerster
We study a class of classification problems best exemplified by the \emph{bank l
oan} problem, where a lender decides whether or not to issue a loan. The lender
only observes whether a customer will repay a loan if the loan is issued to begi
n with, and thus modeled decisions affect what data is available to the lender f
or future decisions. As a result, it is possible for the lender's algorithm to `
`get stuck'' with a self-fulfilling model. This model never corrects its false n
egatives, since it never sees the true label for rejected data, thus accumulatin
g infinite regret. In the case of linear models, this issue can be addressed by
adding optimism directly into the model predictions. However, there are few meth
ods that extend to the function approximation case using Deep Neural Networks. W
e present Pseudo-Label Optimism (PLOT), a conceptually and computationally simpl
e method for this setting applicable to DNNs. \PLOT{} adds an optimistic label t
o the subset of decision points the current model is deciding on, trains the mod
el on all data so far (including these points along with their optimistic labels
), and finally uses the resulting \emph{optimistic} model for decision making. \
PLOT{} achieves competitive performance on a set of three challenging benchmark
problems, requiring minimal hyperparameter tuning. We also show that \PLOT{} sat
isfies a logarithmic regret guarantee, under a Lipschitz and logistic mean label
 model, and under a separability condition on the data.
****
Visualizing the Emergence of Intermediate Visual Patterns in DNNs
Mingjie Li, Shaobo Wang, Quanshi Zhang
This paper proposes a method to visualize the discrimination power of intermedia
te-layer visual patterns encoded by a DNN. Specifically, we visualize (1) how th
e DNN gradually learns regional visual patterns in each intermediate layer durin
g the training process, and (2) the effects of the DNN using non-discriminative
patterns in low layers to construct disciminative patterns in middle/high layers
 through the forward propagation. Based on our visualization method, we can quan
tify knowledge points (i.e. the number of discriminative visual patterns) learne
d by the DNN to evaluate the representation capacity of the DNN. Furthermore, th
is method also provides new insights into signal-processing behaviors of existin
g deep-learning techniques, such as adversarial attacks and knowledge distillati
on.
****
Learning 3D Dense Correspondence via Canonical Point Autoencoder
An-Chieh Cheng, Xueting Li, Min Sun, Ming-Hsuan Yang, Sifei Liu
We propose a canonical point autoencoder (CPAE) that predicts dense corresponden
ces between 3D shapes of the same category. The autoencoder performs two key fun
ctions: (a) encoding an arbitrarily ordered point cloud to a canonical primitive
, e.g., a sphere, and (b) decoding the primitive back to the original input inst
ance shape. As being placed in the bottleneck, this primitive plays a key role t
o map all the unordered point clouds on the canonical surface, and to be reconst
ructed in an ordered fashion. Once trained, points from different shape instance
s that are mapped to the same locations on the primitive surface are determined
to be a pair of correspondence. Our method does not require any form of annotati
on or self-supervised part segmentation network and can handle unaligned input p
oint clouds within a certain rotation range. Experimental results on 3D semantic
 keypoint transfer and part segmentation transfer show that our model performs f
avorably against state-of-the-art correspondence learning methods.
****

Speech-T: Transducer for Text to Speech and Beyond
Jiawei Chen, Xu Tan, Yichong Leng, Jin Xu, Guihua Wen, Tao Qin, Tie-Yan Liu
Neural Transducer (e.g., RNN-T) has been widely used in automatic speech recognition (ASR) due to its capabilities of efficiently modeling monotonic alignments between input and output sequences and naturally supporting streaming inputs. Considering that monotonic alignments are also critical to text to speech (TTS) synthesis and streaming TTS is also an important application scenario, in this work, we explore the possibility of applying Transducer to TTS and more. However, it is challenging because it is difficult to trade off the emission (continuous mel-spectrogram prediction) probability and transition (ASR Transducer predicts blank token to indicate transition to next input) probability when calculating the output probability lattice in Transducer, and it is not easy to learn the alignments between text and speech through the output probability lattice. We propose SpeechTransducer (Speech-T for short), a Transformer based Transducer model that 1) uses a new forward algorithm to separate the transition prediction from the continuous mel-spectrogram prediction when calculating the output probability lattice, and uses a diagonal constraint in the probability lattice to help the alignment learning; 2) supports both full-sentence or streaming TTS by adjusting the look-ahead context; and 3) further supports both TTS and ASR together for the first time, which enjoys several advantages including fewer parameters as well as streaming synthesis and recognition in a single model. Experiments on LJSpeech datasets demonstrate that Speech-T 1) is more robust than the attention based autoregressive TTS model due to its inherent monotonic alignments between text and speech; 2) naturally supports streaming TTS with good voice quality; and 3) enjoys the benefit of joint modeling TTS and ASR in a single network.
****
Multi-modal Dependency Tree for Video Captioning
Wentian Zhao, Xinxiao Wu, Jiebo Luo
Generating fluent and relevant language to describe visual content is critical for the video captioning task. Many existing methods generate captions using sequence models that predict words in a left-to-right order. In this paper, we investigate a graph-structured model for caption generation by explicitly modeling the hierarchical structure in the sentences to further improve the fluency and relevance of sentences. To this end, we propose a novel video captioning method that generates a sentence by first constructing a multi-modal dependency tree and then traversing the constructed tree, where the syntactic structure and semantic relationship in the sentence are represented by the tree topology. To take full advantage of the information from both vision and language, both the visual and textual representation features are encoded into each tree node. Different from existing dependency parsing methods that generate uni-modal dependency trees for language understanding, our method construct s multi-modal dependency trees for language generation of images and videos. We also propose a tree-structured reinforcement learning algorithm to effectively optimize the captioning model where a novel reward is designed by evaluating the semantic consistency between the generated sub-tree and the ground-truth tree. Extensive experiments on several video captioning datasets demonstrate the effectiveness of the proposed method.
****
Greedy and Random Quasi-Newton Methods with Faster Explicit Superlinear Convergence
Dachao Lin, Haishan Ye, Zhihua Zhang
In this paper, we follow Rodomanov and Nesterov’s work to study quasi-Newton methods. We focus on the common SR1 and BFGS quasi-Newton methods to establish better explicit (local) superlinear convergence rates. First, based on the greedy quasi-Newton update which greedily selects the direction to maximize a certain measure of progress, we improve the convergence rate to a condition-number-free superlinear convergence rate. Second, based on the random quasi-Newton update that selects the direction randomly from a spherically symmetric distribution, we show the same superlinear convergence rate established as above. Our analysis is closely related to the approximation of a given Hessian matrix, unconstrained quadratic objective, as well as the general strongly convex, smooth, and strongly se

lf-concordant functions.
****
Neural Tangent Kernel Maximum Mean Discrepancy
Xiuyuan Cheng, Yao Xie
We present a novel neural network Maximum Mean Discrepancy (MMD) statistic by identifying a new connection between neural tangent kernel (NTK) and MMD. This connection enables us to develop a computationally efficient and memory-efficient approach to compute the MMD statistic and perform NTK based two-sample tests towards addressing the long-standing challenge of memory and computational complexity of the MMD statistic, which is essential for online implementation to assimilating new samples. Theoretically, such a connection allows us to understand the NTK test statistic properties, such as the Type-I error and testing power for performing the two-sample test, by adapting existing theories for kernel MMD. Numerical experiments on synthetic and real-world datasets validate the theory and demonstrate the effectiveness of the proposed NTK-MMD statistic.
****
Subgraph Federated Learning with Missing Neighbor Generation
Ke ZHANG, Carl Yang, Xiaoxiao Li, Lichao Sun, Siu Ming Yiu
Graphs have been widely used in data mining and machine learning due to their unique representation of real-world objects and their interactions. As graphs are getting bigger and bigger nowadays, it is common to see their subgraphs separately collected and stored in multiple local systems. Therefore, it is natural to consider the subgraph federated learning setting, where each local system holds a small subgraph that may be biased from the distribution of the whole graph. Hence, the subgraph federated learning aims to collaboratively train a powerful and generalizable graph mining model without directly sharing their graph data. In this work, towards the novel yet realistic setting of subgraph federated learning, we propose two major techniques: (1) FedSage, which trains a GraphSage model based on FedAvg to integrate node features, link structures, and task labels on multiple local subgraphs; (2) FedSage+, which trains a missing neighbor generator along FedSage to deal with missing links across local subgraphs. Empirical results on four real-world graph datasets with synthesized subgraph federated learning settings demonstrate the effectiveness and efficiency of our proposed techniques. At the same time, consistent theoretical implications are made towards their generalization ability on the global graphs.
****
Bellman-consistent Pessimism for Offline Reinforcement Learning
Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, Alekh 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.
****
Can You Learn an Algorithm? Generalizing from Easy to Hard Problems with Recurrent Networks
Avi Schwarzschild, Eitan Borgnia, Arjun Gupta, Furong Huang, Uzi Vishkin, Micah Goldblum, Tom Goldstein
Deep neural networks are powerful machines for visual pattern recognition, but reasoning tasks that are easy for humans may still be difficult for neural models. Humans possess the ability to extrapolate reasoning strategies learned on simple problems to solve harder examples, often by thinking for longer. For example, a person who has learned to solve small mazes can easily extend the very same search techniques to solve much larger mazes by spending more time. In computers, this behavior is often achieved through the use of algorithms, which scale to arbitrarily hard problem instances at the cost of more computation. In contrast, the sequential computing budget of feed-forward neural networks is limited by their depth, and networks trained on simple problems have no way of extending their reasoning to accommodate harder problems. In this work, we show that recurrent networks trained to solve simple problems with few recurrent steps can indeed solve much more complex problems simply by performing additional recurrences dur

ing inference. We demonstrate this algorithmic behavior of recurrent networks on prefix sum computation, mazes, and chess. In all three domains, networks trained on simple problem instances are able to extend their reasoning abilities at test time simply by "thinking for longer."
****
Sub-Linear Memory: How to Make Performers SLiM
Valerii Likhosherstov, Krzysztof M. Choromanski, Jared Quincy Davis, Xingyou Song, Adrian Weller
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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-Continuous Computation Graphs
David Friede, Mathias Niepert
Numerous models for supervised and reinforcement learning benefit from combinations of discrete and continuous model components. End-to-end learnable discrete-continuous models are compositional, tend to generalize better, and are more interpretable. A popular approach to building discrete-continuous computation graphs is that of integrating discrete probability distributions into neural networks using stochastic softmax tricks. Prior work has mainly focused on computation graphs with a single discrete component on each of the graph's execution paths. We analyze the behavior of more complex stochastic computations graphs with multiple sequential discrete components. We show that it is challenging to optimize the parameters of these models, mainly due to small gradients and local minima. We then propose two new strategies to overcome these challenges. First, we show that increasing the scale parameter of the Gumbel noise perturbations during training improves the learning behavior. Second, we propose dropout residual connections specifically tailored to stochastic, discrete-continuous computation graphs. With an extensive set of experiments, we show that we can train complex discrete-continuous models which one cannot train with standard stochastic softmax tricks. We also show that complex discrete-stochastic models generalize better than their continuous counterparts on several benchmark datasets.
****
VQ-GNN: A Universal Framework to Scale up Graph Neural Networks using Vector Quantization
Mucong Ding, Kezhi Kong, Jingling Li, Chen Zhu, John Dickerson, Furong Huang, Tom Goldstein
Most state-of-the-art Graph Neural Networks (GNNs) can be defined as a form of graph convolution which can be realized by message passing between direct neighbors or beyond. To scale such GNNs to large graphs, various neighbor-, layer-, or subgraph-sampling techniques are proposed to alleviate the "neighbor explosion" problem by considering only a small subset of messages passed to the nodes in a mini-batch. However, sampling-based methods are difficult to apply to GNNs that utilize many-hops-away or global context each layer, show unstable performance for different tasks and datasets, and do not speed up model inference. We propose a principled and fundamentally different approach, VQ-GNN, a universal framework to scale up any convolution-based GNNs using Vector Quantization (VQ) without compromising the performance. In contrast to sampling-based techniques, our approach can effectively preserve all the messages passed to a mini-batch of nodes by learning and updating a small number of quantized reference vectors of global node representations, using VQ within each GNN layer. Our framework avoids the "neighbor explosion" problem of GNNs using quantized representations combined with a low-rank version of the graph convolution matrix. We show that such a compact low-rank version of the gigantic convolution matrix is sufficient both theoretically and experimentally. In company with VQ, we design a novel approximated message passing algorithm and a nontrivial back-propagation rule for our framework. Experiments on various types of GNN backbones demonstrate the scalability and competitive performance of our framework on large-graph node classification and link prediction benchmarks.

****
Overcoming Catastrophic Forgetting in Incremental Few-Shot Learning by Finding F
lat Minima
Guangyuan SHI, JIAXIN CHEN, Wenlong Zhang, Li-Ming Zhan, Xiao-Ming Wu
This paper considers incremental few-shot learning, which requires a model to co
ntinually recognize new categories with only a few examples provided. Our study
shows that existing methods severely suffer from catastrophic forgetting, a well
-known problem in incremental learning, which is aggravated due to data scarcity
 and imbalance in the few-shot setting. Our analysis further suggests that to pr
event catastrophic forgetting, actions need to be taken in the primitive stage -
- the training of base classes instead of later few-shot learning sessions. Ther
efore, we propose to search for flat local minima of the base training objective
 function and then fine-tune the model parameters within the flat region on new
tasks. In this way, the model can efficiently learn new classes while preserving
 the old ones. Comprehensive experimental results demonstrate that our approach
outperforms all prior state-of-the-art methods and is very close to the approxim
ate upper bound. The source code is available at https://github.com/moukamisama/
F2M.
****
Functional Neural Networks for Parametric Image Restoration Problems
Fangzhou Luo, Xiaolin Wu, Yanhui Guo
Almost every single image restoration problem has a closely related parameter, s
uch as the scale factor in super-resolution, the noise level in image denoising,
 and the quality factor in JPEG deblocking. Although recent studies on image res
toration problems have achieved great success due to the development of deep neu
ral networks, they handle the parameter involved in an unsophisticated way. Most
 previous researchers either treat problems with different parameter levels as i
ndependent tasks, and train a specific model for each parameter level; or simply
 ignore the parameter, and train a single model for all parameter levels. The tw
o popular approaches have their own shortcomings. The former is inefficient in c
omputing and the latter is ineffective in performance. In this work, we propose
a novel system called functional neural network (FuncNet) to solve a parametric
image restoration problem with a single model. Unlike a plain neural network, th
e smallest conceptual element of our FuncNet is no longer a floating-point varia
ble, but a function of the parameter of the problem. This feature makes it both
efficient and effective for a parametric problem. We apply FuncNet to super-reso
lution, image denoising, and JPEG deblocking. The experimental results show the
superiority of our FuncNet on all three parametric image restoration tasks over
the state of the arts.
****
Intrinsic Dimension, Persistent Homology and Generalization in Neural Networks
Tolga Birdal, Aaron Lou, Leonidas J. Guibas, Umut Simsekli
Disobeying the classical wisdom of statistical learning theory, modern deep neur
al networks generalize well even though they typically contain millions of param
eters. Recently, it has been shown that the trajectories of iterative optimizati
on algorithms can possess \emph{fractal structures}, and their generalization er
ror can be formally linked to the complexity of such fractals. This complexity i
s measured by the fractal's \emph{intrinsic dimension}, a quantity usually much
smaller than the number of parameters in the network. Even though this perspecti
ve provides an explanation for why overparametrized networks would not overfit,
computing the intrinsic dimension (\eg, for monitoring generalization during tra
ining) is a notoriously difficult task, where existing methods typically fail e
ven in moderate ambient dimensions. In this study, we consider this problem from
 the lens of topological data analysis (TDA) and develop a generic computational
 tool that is built on rigorous mathematical foundations. By making a novel conn
ection between learning theory and TDA, we first illustrate that the generalizat
ion error can be equivalently bounded in terms of a notion called the 'persisten
t homology dimension' (PHD), where, compared with prior work, our approach does
not require any additional geometrical or statistical assumptions on the trainin
g dynamics. Then, by utilizing recently established theoretical results and TDA

tools, we develop an efficient algorithm to estimate PHD in the scale of modern deep neural networks and further provide visualization tools to help understand generalization in deep learning. Our experiments show that the proposed approach can efficiently compute a network's intrinsic dimension in a variety of settings, which is predictive of the generalization error.
****

GemNet: Universal Directional Graph Neural Networks for Molecules

Johannes Gasteiger, Florian Becker, Stephan Günnemann

Effectively predicting molecular interactions has the potential to accelerate molecular dynamics by multiple orders of magnitude and thus revolutionize chemical simulations. Graph neural networks (GNNs) have recently shown great successes for this task, overtaking classical methods based on fixed molecular kernels. However, they still appear very limited from a theoretical perspective, since regular GNNs cannot distinguish certain types of graphs. In this work we close this gap between theory and practice. We show that GNNs with directed edge embeddings and two-hop message passing are indeed universal approximators for predictions that are invariant to translation, and equivariant to permutation and rotation. We then leverage these insights and multiple structural improvements to propose the geometric message passing neural network (GemNet). We demonstrate the benefits of the proposed changes in multiple ablation studies. GemNet outperforms previous models on the COLL, MD17, and OC20 datasets by 34%, 41%, and 20%, respectively, and performs especially well on the most challenging molecules. Our implementation is available online.
****

Loss function based second-order Jensen inequality and its application to particle variational inference

Futoshi Futami, Tomoharu Iwata, naonori ueda, Issei Sato, Masashi Sugiyama

Bayesian model averaging, obtained as the expectation of a likelihood function by a posterior distribution, has been widely used for prediction, evaluation of uncertainty, and model selection. Various approaches have been developed to efficiently capture the information in the posterior distribution; one such approach is the optimization of a set of models simultaneously with interaction to ensure the diversity of the individual models in the same way as ensemble learning. A representative approach is particle variational inference (PVI), which uses an ensemble of models as an empirical approximation for the posterior distribution. PVI iteratively updates each model with a repulsion force to ensure the diversity of the optimized models. However, despite its promising performance, a theoretical understanding of this repulsion and its association with the generalization ability remains unclear. In this paper, we tackle this problem in light of PAC-Bayesian analysis. First, we provide a new second-order Jensen inequality, which has the repulsion term based on the loss function. Thanks to the repulsion term, it is tighter than the standard Jensen inequality. Then, we derive a novel generalization error bound and show that it can be reduced by enhancing the diversity of models. Finally, we derive a new PVI that optimizes the generalization error bound directly. Numerical experiments demonstrate that the performance of the proposed PVI compares favorably with existing methods in the experiment.
****

Detecting and Adapting to Irregular Distribution Shifts in Bayesian Online Learning

Aodong Li, Alex Boyd, Padhraic Smyth, Stephan Mandt

We consider the problem of online learning in the presence of distribution shifts that occur at an unknown rate and of unknown intensity. We derive a new Bayesian online inference approach to simultaneously infer these distribution shifts and adapt the model to the detected changes by integrating ideas from change point detection, switching dynamical systems, and Bayesian online learning. Using a binary ‘change variable,’ we construct an informative prior such that--if a change is detected--the model partially erases the information of past model updates by tempering to facilitate adaptation to the new data distribution. Furthermore, the approach uses beam search to track multiple change-point hypotheses and selects the most probable one in hindsight. Our proposed method is model-agnostic,

applicable in both supervised and unsupervised learning settings, suitable for an environment of concept drifts or covariate drifts, and yields improvements over state-of-the-art Bayesian online learning approaches.
****
Asynchronous Decentralized SGD with Quantized and Local Updates
Giorgi Nadiradze, Amirmojtaba Sabour, Peter Davies, Shigang Li, Dan Alistarh
Decentralized optimization is emerging as a viable alternative for scalable distributed machine learning, but also introduces new challenges in terms of synchronization costs. To this end, several communication-reduction techniques, such as non-blocking communication, quantization, and local steps, have been explored in the decentralized setting. Due to the complexity of analyzing optimization in such a relaxed setting, this line of work often assumes \emph{global} communication rounds, which require additional synchronization. In this paper, we consider decentralized optimization in the simpler, but harder to analyze, \emph{asynchronous gossip} model, in which communication occurs in discrete, randomly chosen pairings among nodes. Perhaps surprisingly, we show that a variant of SGD called \emph{SwarmSGD} still converges in this setting, even if \emph{non-blocking communication}, \emph{quantization}, and \emph{local steps} are all applied \emph{in conjunction}, and even if the node data distributions and underlying graph topology are both \emph{heterogenous}. Our analysis is based on a new connection with multi-dimensional load-balancing processes. We implement this algorithm and deploy it in a super-computing environment, showing that it can outperform previous decentralized methods in terms of end-to-end training time, and that it can even rival carefully-tuned large-batch SGD for certain tasks.
****
Stochastic Shortest Path: Minimax, Parameter-Free and Towards Horizon-Free Regret
Jean Tarbouriech, Runlong Zhou, Simon S. Du, 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-authors prior to requesting a name change in the electronic proceedings.
****
Nested Counterfactual Identification from Arbitrary Surrogate Experiments
Juan Correa, Sanghack Lee, Elias Bareinboim
The Ladder of Causation describes three qualitatively different types of activities an agent may be interested in engaging in, namely, seeing (observational), doing (interventional), and imagining (counterfactual) (Pearl and Mackenzie, 2018). The inferential challenge imposed by the causal hierarchy is that data is collected by an agent observing or intervening in a system (layers 1 and 2), while its goal may be to understand what would have happened had it taken a different course of action, contrary to what factually ended up happening (layer 3). While there exists a solid understanding of the conditions under which cross-layer inferences are allowed from observations to interventions, the results are somewhat scarcer when targeting counterfactual quantities. In this paper, we study the identification of nested counterfactuals from an arbitrary combination of observations and experiments. Specifically, building on a more explicit definition of nested counterfactuals, we prove the counterfactual unnesting theorem (CUT), which allows one to map arbitrary nested counterfactuals to unnested ones. For instance, applications in mediation and fairness analysis usually evoke notions of direct, indirect, and spurious effects, which naturally require nesting. Second, we introduce a sufficient and necessary graphical condition for counterfactual identification from an arbitrary combination of observational and experimental distributions. Lastly, we develop an efficient and complete algorithm for identifying nested counterfactuals; failure of the algorithm returning an expression for a query implies it is not identifiable.
****
Sim and Real: Better Together
Shirli Di-Castro, Dotan Di Castro, Shie Mannor

Simulation is used extensively in autonomous systems, particularly in robotic ma nipulation. By far, the most common approach is to train a controller in simulat ion, and then use it as an initial starting point for the real system. We demons trate how to learn simultaneously from both simulation and interaction with the real environment. We propose an algorithm for balancing the large number of samp les from the high throughput but less accurate simulation and the low-throughput , high-fidelity and costly samples from the real environment. We achieve that by maintaining a replay buffer for each environment the agent interacts with. We a nalyze such multi-environment interaction theoretically, and provide convergence properties, through a novel theoretical replay buffer analysis. We demonstrate the efficacy of our method on a sim-to-real environment.
****
Trustworthy Multimodal Regression with Mixture of Normal-inverse Gamma Distribut ions
Huan Ma, Zongbo Han, Changqing Zhang, Huazhu Fu, Joey Tianyi Zhou, Qinghua Hu
Multimodal regression is a fundamental task, which integrates the information fr om different sources to improve the performance of follow-up applications. Howev er, existing methods mainly focus on improving the performance and often ignore the confidence of prediction for diverse situations. In this study, we are devot ed to trustworthy multimodal regression which is critical in cost-sensitive doma ins. To this end, we introduce a novel Mixture of Normal-Inverse Gamma distribut ions (MoNIG) algorithm, which efficiently estimates uncertainty in principle for adaptive integration of different modalities and produces a trustworthy regress ion result. Our model can be dynamically aware of uncertainty for each modality, and also robust for corrupted modalities. Furthermore, the proposed MoNIG ensur es explicitly representation of (modality-specific/global) epistemic and aleator ic uncertainties, respectively. Experimental results on both synthetic and diffe rent real-world data demonstrate the effectiveness and trustworthiness of our me thod on various multimodal regression tasks (e.g., temperature prediction for su perconductivity, relative location prediction for CT slices, and multimodal sent iment analysis).
****
An Empirical Study of Adder Neural Networks for Object Detection
Xinghao Chen, Chang Xu, Minjing Dong, 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-auth ors prior to requesting a name change in the electronic proceedings.
****
Does Knowledge Distillation Really Work?
Samuel Stanton, Pavel Izmailov, Polina Kirichenko, Alexander A. Alemi, Andrew G. Wilson
Knowledge distillation is a popular technique for training a small student netwo rk to emulate a larger teacher model, such as an ensemble of networks. We show t hat while knowledge distillation can improve student generalization, it does not typically work as it is commonly understood: there often remains a surprisingly large discrepancy between the predictive distributions of the teacher and the s tudent, even in cases when the student has the capacity to perfectly match the t eacher. We identify difficulties in optimization as a key reason for why the stu dent is unable to match the teacher. We also show how the details of the dataset used for distillation play a role in how closely the student matches the teache r --- and that more closely matching the teacher paradoxically does not always l ead to better student generalization.
****
Teachable Reinforcement Learning via Advice Distillation
Olivia Watkins, Abhishek Gupta, Trevor Darrell, Pieter Abbeel, Jacob Andreas
Training automated agents to complete complex tasks in interactive environments is challenging: reinforcement learning requires careful hand-engineering of rewa rd functions, imitation learning requires specialized infrastructure and access to a human expert, and learning from intermediate forms of supervision (like bin

ary preferences) is time-consuming and extracts little information from each human intervention. Can we overcome these challenges by building agents that learn from rich, interactive feedback instead? We propose a new supervision paradigm for interactive learning based on "teachable" decision-making systems that learn from structured advice provided by an external teacher. We begin by formalizing a class of human-in-the-loop decision making problems in which multiple forms of teacher-provided advice are available to a learner. We then describe a simple learning algorithm for these problems that first learns to interpret advice, then learns from advice to complete tasks even in the absence of human supervision. In puzzle-solving, navigation, and locomotion domains, we show that agents that learn from advice can acquire new skills with significantly less human supervision than standard reinforcement learning algorithms and often less than imitation learning.
****
Antipodes of Label Differential Privacy: PATE and ALIBI
Mani Malek Esmaeili, Ilya Mironov, Karthik Prasad, Igor Shilov, Florian Tramer
We consider the privacy-preserving machine learning (ML) setting where the trained model must satisfy differential privacy (DP) with respect to the labels of the training examples. We propose two novel approaches based on, respectively, the Laplace mechanism and the PATE framework, and demonstrate their effectiveness on standard benchmarks.While recent work by Ghazi et al. proposed Label DP schemes based on a randomized response mechanism, we argue that additive Laplace noise coupled with Bayesian inference (ALIBI) is a better fit for typical ML tasks. Moreover, we show how to achieve very strong privacy levels in some regimes, with our adaptation of the PATE framework that builds on recent advances in semi-supervised learning.We complement theoretical analysis of our algorithms' privacy guarantees with empirical evaluation of their memorization properties. Our evaluation suggests that comparing different algorithms according to their provable DP guarantees can be misleading and favor a less private algorithm with a tighter analysis.Code for implementation of algorithms and memorization attacks is available from https://github.com/facebookresearch/labeldpantipodes.
****
Visual Search Asymmetry: Deep Nets and Humans Share Similar Inherent Biases
Shashi Kant Gupta, Mengmi Zhang, CHIA-CHIEN WU, Jeremy Wolfe, Gabriel Kreiman
Visual search is a ubiquitous and often challenging daily task, exemplified by looking for the car keys at home or a friend in a crowd. An intriguing property of some classical search tasks is an asymmetry such that finding a target A among distractors B can be easier than finding B among A. To elucidate the mechanisms responsible for asymmetry in visual search, we propose a computational model that takes a target and a search image as inputs and produces a sequence of eye movements until the target is found. The model integrates eccentricity-dependent visual recognition with target-dependent top-down cues. We compared the model against human behavior in six paradigmatic search tasks that show asymmetry in humans. Without prior exposure to the stimuli or task-specific training, the model provides a plausible mechanism for search asymmetry. We hypothesized that the polarity of search asymmetry arises from experience with the natural environment. We tested this hypothesis by training the model on augmented versions of ImageNet where the biases of natural images were either removed or reversed. The polarity of search asymmetry disappeared or was altered depending on the training protocol. This study highlights how classical perceptual properties can emerge in neural network models, without the need for task-specific training, but rather as a consequence of the statistical properties of the developmental diet fed to the model. All source code and data are publicly available at https://github.com/kreimanlab/VisualSearchAsymmetry.
****
On the Universality of Graph Neural Networks on Large Random Graphs
Nicolas Keriven, Alberto Bietti, Samuel Vaiter
We study the approximation power of Graph Neural Networks (GNNs) on latent position random graphs. In the large graph limit, GNNs are known to converge to certain ``continuous'' models known as c-GNNs, which directly enables a study of thei

r approximation power on random graph models. In the absence of input node featu
res however, just as GNNs are limited by the Weisfeiler-Lehman isomorphism test,
 c-GNNs will be severely limited on simple random graph models. For instance, th
ey will fail to distinguish the communities of a well-separated Stochastic Block
 Model (SBM) with constant degree function. Thus, we consider recently proposed
architectures that augment GNNs with unique node identifiers, referred to as Str
uctural GNNs here (SGNNs). We study the convergence of SGNNs to their continuous
 counterpart (c-SGNNs) in the large random graph limit, under new conditions on
the node identifiers. We then show that c-SGNNs are strictly more powerful than
c-GNNs in the continuous limit, and prove their universality on several random g
raph models of interest, including most SBMs and a large class of random geometr
ic graphs. Our results cover both permutation-invariant and permutation-equivari
ant architectures.
****
Inverse Reinforcement Learning in a Continuous State Space with Formal Guarantee
s
Gregory Dexter, Kevin Bello, Jean Honorio
Inverse Reinforcement Learning (IRL) is the problem of finding a reward function
 which describes observed/known expert behavior. The IRL setting is remarkably
useful for automated control, in situations where the reward function is difficu
lt to specify manually or as a means to extract agent preference. In this work,
we provide a new IRL algorithm for the continuous state space setting with unkno
wn transition dynamics by modeling the system using a basis of orthonormal funct
ions. Moreover, we provide a proof of correctness and formal guarantees on the s
ample and time complexity of our algorithm. Finally, we present synthetic exper
iments to corroborate our theoretical guarantees.
****
Adversarial Attacks on Graph Classifiers via Bayesian Optimisation
Xingchen Wan, Henry Kenlay, Robin Ru, Arno Blaas, Michael A Osborne, Xiaowen Don
g
Graph neural networks, a popular class of models effective in a wide range of gr
aph-based learning tasks, have been shown to be vulnerable to adversarial attack
s. While the majority of the literature focuses on such vulnerability in node-le
vel classification tasks, little effort has been dedicated to analysing adversar
ial attacks on graph-level classification, an important problem with numerous re
al-life applications such as biochemistry and social network analysis. The few e
xisting methods often require unrealistic setups, such as access to internal inf
ormation of the victim models, or an impractically-large number of queries. We p
resent a novel Bayesian optimisation-based attack method for graph classificatio
n models. Our method is black-box, query-efficient and parsimonious with respect
 to the perturbation applied. We empirically validate the effectiveness and flex
ibility of the proposed method on a wide range of graph classification tasks inv
olving varying graph properties, constraints and modes of attack. Finally, we an
alyse common interpretable patterns behind the adversarial samples produced, whi
ch may shed further light on the adversarial robustness of graph classification
models.
****
Regulating algorithmic filtering on social media
Sarah Cen, Devavrat Shah
By filtering the content that users see, social media platforms have the ability
 to influence users' perceptions and decisions, from their dining choices to the
ir voting preferences. This influence has drawn scrutiny, with many calling for
regulations on filtering algorithms, but designing and enforcing regulations rem
ains challenging. In this work, we examine three questions. First, given a regul
ation, how would one design an audit to enforce it? Second, does the audit impos
e a performance cost on the platform? Third, how does the audit affect the conte
nt that the platform is incentivized to filter? In response to these questions,
we propose a method such that, given a regulation, an auditor can test whether t
hat regulation is met with only black-box access to the filtering algorithm. We
then turn to the platform's perspective. The platform's goal is to maximize an o

bjective function while meeting regulation. We find that there are conditions un
der which the regulation does not place a high performance cost on the platform
and, notably, that content diversity can play a key role in aligning the interes
ts of the platform and regulators.
****
argmax centroid
Chengyue Gong, Mao Ye, Qiang Liu
We propose a general method to construct centroid approximation for the distribu
tion of maximum points of a random function (a.k.a. argmax distribution), which
finds broad applications in machine learning. Our method optimizes a set of cent
roid points to compactly approximate the argmax distribution with a simple objec
tive function, without explicitly drawing exact samples from the argmax distribu
tion. Theoretically, the argmax centroid method can be shown to minimize a surro
gate of Wasserstein distance between the ground-truth argmax distribution and th
e centroid approximation under proper conditions. We demonstrate the applicabili
ty and effectiveness of our method on a variety of real-world multi-task learnin
g applications, including few-shot image classification, personalized dialogue s
ystems and multi-target domain adaptation.
****
Contrastive Learning of Global and Local Video Representations
shuang ma, Zhaoyang Zeng, Daniel McDuff, Yale Song
Contrastive learning has delivered impressive results for various tasks in the s
elf-supervised regime. However, existing approaches optimize for learning repres
entations specific to downstream scenarios, i.e., global representations suitabl
e for tasks such as classification or local representations for tasks such as de
tection and localization. While they produce satisfactory results in the intende
d downstream scenarios, they often fail to generalize to tasks that they were no
t originally designed for. In this work, we propose to learn video representatio
ns that generalize to both the tasks which require global semantic information (
e.g., classification) and the tasks that require local fine-grained spatio-tempo
ral information (e.g., localization). We achieve this by optimizing two contrast
ive objectives that together encourage our model to learn global-local visual in
formation given audio signals. We show that the two objectives mutually improve
the generalizability of the learned global-local representations, significantly
outperforming their disjointly learned counterparts. We demonstrate our approach
on various tasks including action/sound classification, lipreading, deepfake de
tection, event and sound localization.
****
BooVI: Provably Efficient Bootstrapped Value Iteration
Boyi Liu, Qi Cai, Zhuoran Yang, Zhaoran Wang
Requests for name changes in the electronic proceedings will be accepted with no
questions asked. However name changes may cause bibliographic tracking issues.
Authors are asked to consider this carefully and discuss it with their co-auth
ors prior to requesting a name change in the electronic proceedings.
****
Do Wider Neural Networks Really Help Adversarial Robustness?
Boxi Wu, Jinghui Chen, Deng Cai, Xiaofei He, Quanquan Gu
Adversarial training is a powerful type of defense against adversarial examples.
Previous empirical results suggest that adversarial training requires wider net
works for better performances. However, it remains elusive how does neural netwo
rk width affect model robustness. In this paper, we carefully examine the relati
onship between network width and model robustness. Specifically, we show that th
e model robustness is closely related to the tradeoff between natural accuracy a
nd perturbation stability, which is controlled by the robust regularization para
meter λ. With the same λ, wider networks can achieve better natural accuracy but
worse perturbation stability, leading to a potentially worse overall model robu
stness. To understand the origin of this phenomenon, we further relate the pertu
rbation stability with the network's local Lipschitzness. By leveraging recent r
esults on neural tangent kernels, we theoretically show that wider networks tend
to have worse perturbation stability. Our analyses suggest that: 1) the common

strategy of first fine-tuning λ on small networks and then directly use it for w
ide model training could lead to deteriorated model robustness; 2) one needs to
properly enlarge λ to unleash the robustness potential of wider models fully. Fi
nally, we propose a new Width Adjusted Regularization (WAR) method that adaptive
ly enlarges λ on wide models and significantly saves the tuning time.
****
Exploring the Limits of Out-of-Distribution Detection
Stanislav Fort, Jie Ren, Balaji Lakshminarayanan
Near out-of-distribution detection (OOD) is a major challenge for deep neural ne
tworks. We demonstrate that large-scale pre-trained transformers can significant
ly improve the state-of-the-art (SOTA) on a range of near OOD tasks across diffe
rent data modalities. For instance, on CIFAR-100 vs CIFAR-10 OOD detection, we i
mprove the AUROC from 85% (current SOTA) to more than 96% using Vision Transform
ers pre-trained on ImageNet21k. On a challenging genomics OOD detection benchmar
k, we improve the AUROC from 66% to 77% using transformer and unsupervised pre-t
raining. To further improve performance, we explore the few-shot outlier exposu
re setting where a few examples from outlier classes may be available; we show t
hat pre-trained transformers are particularly well-suited for outlier exposure,
 and that the AUROC of OOD detection on CIFAR-100 vs CIFAR-10 can be improved t
o 98.7% with just 1 image per OOD class, and 99.46% with 10 images per OOD class
. For multi-modal image-text pre-trained transformers such as CLIP, we explore a
 new way of using just the names of outlier classes as a sole source of informat
ion without any accompanying images, and show that this outperforms previous SOT
A on standard OOD benchmark tasks.
****
ABC: Auxiliary Balanced Classifier for Class-imbalanced Semi-supervised Learning
Hyuck Lee, Seungjae Shin, Heeyoung Kim
Existing semi-supervised learning (SSL) algorithms typically assume class-balanc
ed datasets, although the class distributions of many real world datasets are im
balanced. In general, classifiers trained on a class-imbalanced dataset are bias
ed toward the majority classes. This issue becomes more problematic for SSL algo
rithms because they utilize the biased prediction of unlabeled data for training
. However, traditional class-imbalanced learning techniques, which are designed
for labeled data, cannot be readily combined with SSL algorithms. We propose a s
calable class-imbalanced SSL algorithm that can effectively use unlabeled data,
while mitigating class imbalance by introducing an auxiliary balanced classifier
 (ABC) of a single layer, which is attached to a representation layer of an exis
ting SSL algorithm. The ABC is trained with a class-balanced loss of a minibatch
, while using high-quality representations learned from all data points in the m
inibatch using the backbone SSL algorithm to avoid overfitting and information l
oss. Moreover, we use consistency regularization, a recent SSL technique for uti
lizing unlabeled data in a modified way, to train the ABC to be balanced among t
he classes by selecting unlabeled data with the same probability for each class.
 The proposed algorithm achieves state-of-the-art performance in various class-i
mbalanced SSL experiments using four benchmark datasets.
****
BCD Nets: Scalable Variational Approaches for Bayesian Causal Discovery
Chris Cundy, Aditya Grover, Stefano Ermon
A structural equation model (SEM) is an effective framework to reason over causa
l relationships represented via a directed acyclic graph (DAG).Recent advances h
ave enabled effective maximum-likelihood point estimation of DAGs from observati
onal data. However, a point estimate may not accurately capture the uncertainty
in inferring the underlying graph in practical scenarios, wherein the true DAG i
s non-identifiable and/or the observed dataset is limited.We propose Bayesian Ca
usal Discovery Nets (BCD Nets), a variational inference framework for estimating
 a distribution over DAGs characterizing a linear-Gaussian SEM.Developing a full
 Bayesian posterior over DAGs is challenging due to the the discrete and combina
torial nature of graphs.We analyse key design choices for scalable VI over DAGs,
 such as 1) the parametrization of DAGs via an expressive variational family, 2)
 a continuous relaxation that enables low-variance stochastic optimization, and

3) suitable priors over the latent variables.We provide a series of experiments on real and synthetic data showing that BCD Nets outperform maximum-likelihood methods on standard causal discovery metrics such as structural Hamming distance in low data regimes.
****

Discovering Dynamic Salient Regions for Spatio-Temporal Graph Neural Networks
Iulia Duta, Andrei Nicolicioiu, Marius Leordeanu
Graph Neural Networks are perfectly suited to capture latent interactions between various entities in the spatio-temporal domain (e.g. videos). However, when an explicit structure is not available, it is not obvious what atomic elements should be represented as nodes. Current works generally use pre-trained object detectors or fixed, predefined regions to extract graph nodes. Improving upon this, our proposed model learns nodes that dynamically attach to well-delimited salient regions, which are relevant for a higher-level task, without using any object-level supervision. Constructing these localized, adaptive nodes gives our model inductive bias towards object-centric representations and we show that it discovers regions that are well correlated with objects in the video. In extensive ablation studies and experiments on two challenging datasets, we show superior performance to previous graph neural networks models for video classification.
****

Information-constrained optimization: can adaptive processing of gradients help?
Jayadev Acharya, Clement Canonne, Prathamesh Mayekar, Himanshu Tyagi
We revisit first-order optimization under local information constraints such as local privacy, gradient quantization, and computational constraints limiting access to a few coordinates of the gradient. In this setting, the optimization algorithm is not allowed to directly access the complete output of the gradient oracle, but only gets limited information about it subject to the local information constraints.   We study the role of adaptivity in processing the gradient output to obtain this limited information from it, and obtain tight or nearly tight bounds for both convex and strongly convex optimization when adaptive gradient processing is allowed.
****

Towards Calibrated Model for Long-Tailed Visual Recognition from Prior Perspective
Zhengzhuo Xu, Zenghao Chai, Chun Yuan
Real-world data universally confronts a severe class-imbalance problem and exhibits a long-tailed distribution, i.e., most labels are associated with limited instances. The naïve models supervised by such datasets would prefer dominant labels, encounter a serious generalization challenge and become poorly calibrated. We propose two novel methods from the prior perspective to alleviate this dilemma. First, we deduce a balance-oriented data augmentation named Uniform Mixup (UniMix) to promote mixup in long-tailed scenarios, which adopts advanced mixing factor and sampler in favor of the minority. Second, motivated by the Bayesian theory, we figure out the Bayes Bias (Bayias), an inherent bias caused by the inconsistency of prior, and compensate it as a modification on standard cross-entropy loss. We further prove that both the proposed methods ensure the classification calibration theoretically and empirically. Extensive experiments verify that our strategies contribute to a better-calibrated model, and their combination achieves state-of-the-art performance on CIFAR-LT, ImageNet-LT, and iNaturalist 2018.
****

Learning to Draw: Emergent Communication through Sketching
Daniela Mihai, Jonathon Hare
Evidence that visual communication preceded written language and provided a basis for it goes back to prehistory, in forms such as cave and rock paintings depicting traces of our distant ancestors. Emergent communication research has sought to explore how agents can learn to communicate in order to collaboratively solve tasks. Existing research has focused on language, with a learned communication channel transmitting sequences of discrete tokens between the agents. In this work, we explore a visual communication channel between agents that are allowed to draw with simple strokes. Our agents are parameterised by deep neural networks

, and the drawing procedure is differentiable, allowing for end-to-end training. In the framework of a referential communication game, we demonstrate that agents can not only successfully learn to communicate by drawing, but with appropriate inductive biases, can do so in a fashion that humans can interpret. We hope to encourage future research to consider visual communication as a more flexible and directly interpretable alternative of training collaborative agents.
****
Self-Supervised Learning of Event-Based Optical Flow with Spiking Neural Networks
Jesse Hagenaars, Federico Paredes-Valles, Guido de Croon
The field of neuromorphic computing promises extremely low-power and low-latency sensing and processing. Challenges in transferring learning algorithms from traditional artificial neural networks (ANNs) to spiking neural networks (SNNs) have so far prevented their application to large-scale, complex regression tasks. Furthermore, realizing a truly asynchronous and fully neuromorphic pipeline that maximally attains the abovementioned benefits involves rethinking the way in which this pipeline takes in and accumulates information. In the case of perception, spikes would be passed as-is and one-by-one between an event camera and an SNN, meaning all temporal integration of information must happen inside the network. In this article, we tackle these two problems. We focus on the complex task of learning to estimate optical flow from event-based camera inputs in a self-supervised manner, and modify the state-of-the-art ANN training pipeline to encode minimal temporal information in its inputs. Moreover, we reformulate the self-supervised loss function for event-based optical flow to improve its convexity. We perform experiments with various types of recurrent ANNs and SNNs using the proposed pipeline. Concerning SNNs, we investigate the effects of elements such as parameter initialization and optimization, surrogate gradient shape, and adaptive neuronal mechanisms. We find that initialization and surrogate gradient width play a crucial part in enabling learning with sparse inputs, while the inclusion of adaptivity and learnable neuronal parameters can improve performance. We show that the performance of the proposed ANNs and SNNs are on par with that of the current state-of-the-art ANNs trained in a self-supervised manner.
****
On the Value of Infinite Gradients in Variational Autoencoder Models
Bin Dai, Li Wenliang, David Wipf
A number of recent studies of continuous variational autoencoder (VAE) models have noted, either directly or indirectly, the tendency of various parameter gradients to drift towards infinity during training. Because such gradients could potentially contribute to numerical instabilities, and are often framed as a problematic phenomena to be avoided, it may be tempting to shift to alternative energy functions that guarantee bounded gradients. But it remains an open question: What might the unintended consequences of such a restriction be? To address this issue, we examine how unbounded gradients relate to the regularization of a broad class of autoencoder-based architectures, including VAE models, as applied to data lying on or near a low-dimensional manifold (e.g., natural images). Our main finding is that, if the ultimate goal is to simultaneously avoid over-regularization (high reconstruction errors, sometimes referred to as posterior collapse) and under-regularization (excessive latent dimensions are not pruned from the model), then an autoencoder-based energy function with infinite gradients around optimal representations is provably required per a certain technical sense which we carefully detail. Given that both over- and under-regularization can directly lead to poor generated sample quality or suboptimal feature selection, this result suggests that heuristic modifications to or constraints on the VAE energy function may at times be ill-advised, and large gradients should be accommodated to the extent possible.
****
Online Robust Reinforcement Learning with Model Uncertainty
Yue Wang, Shaofeng Zou
Robust reinforcement learning (RL) is to find a policy that optimizes the worst-case performance over an uncertainty set of MDPs. In this paper, we focus on mod

el-free robust RL, where the uncertainty set is defined to be centering at a mis specified MDP that generates samples, and is assumed to be unknown. We develop a sample-based approach to estimate the unknown uncertainty set, and design robust Q-learning algorithm (tabular case) and robust TDC algorithm (function approximation setting), which can be implemented in an online and incremental fashion. For the robust Q-learning algorithm, we prove that it converges to the optimal robust Q function, and for the robust TDC algorithm, we prove that it converges asymptotically to some stationary points. Unlike the results in [Roy et al., 2017], our algorithms do not need any additional conditions on the discount factor to guarantee the convergence. We further characterize the finite-time error bounds of the two algorithms, and show that both the robust Q-learning and robust TDC algorithms converge as fast as their vanilla counterparts (within a constant factor). Our numerical experiments further demonstrate the robustness of our algorithms. Our approach can be readily extended to robustify many other algorithms, e.g., TD, SARSA, and other GTD algorithms.
****

Neural View Synthesis and Matching for Semi-Supervised Few-Shot Learning of 3D Pose

Angtian Wang, Shenxiao Mei, Alan L. Yuille, Adam Kortylewski

We study the problem of learning to estimate the 3D object pose from a few labelled examples and a collection of unlabelled data. Our main contribution is a learning framework, neural view synthesis and matching, that can transfer the 3D pose annotation from the labelled to unlabelled images reliably, despite unseen 3D views and nuisance variations such as the object shape, texture, illumination or scene context. In our approach, objects are represented as 3D cuboid meshes composed of feature vectors at each mesh vertex. The model is initialized from a few labelled images and is subsequently used to synthesize feature representations of unseen 3D views. The synthesized views are matched with the feature representations of unlabelled images to generate pseudo-labels of the 3D pose. The pseudo-labelled data is, in turn, used to train the feature extractor such that the features at each mesh vertex are more invariant across varying 3D views of the object. Our model is trained in an EM-type manner alternating between increasing the 3D pose invariance of the feature extractor and annotating unlabelled data through neural view synthesis and matching. We demonstrate the effectiveness of the proposed semi-supervised learning framework for 3D pose estimation on the PASCAL3D+ and KITTI datasets. We find that our approach outperforms all baselines by a wide margin, particularly in an extreme few-shot setting where only 7 annotated images are given. Remarkably, we observe that our model also achieves an exceptional robustness in out-of-distribution scenarios that involve partial occlusion.
****

Sharp Impossibility Results for Hyper-graph Testing

Jiashun Jin, Zheng Tracy Ke, Jiajun Liang

In a broad Degree-Corrected Mixed-Membership (DCMM) setting, we test whether a non-uniform hypergraph has only one community or has multiple communities. Since both the null and alternative hypotheses have many unknown parameters, the challenge is, given an alternative, how to identify the null that is hardest to separate from the alternative. We approach this by proposing a degree matching strategy where the main idea is leveraging the theory for tensor scaling to create a least favorable pair of hypotheses. We present a result on standard minimax lower bound theory and a result on Region of Impossibility (which is more informative than the minimax lower bound). We show that our lower bounds are tight by introducing a new test that attains the lower bound up to a logarithmic factor. We also discuss the case where the hypergraphs may have mixed-memberships.
****

Evaluating Gradient Inversion Attacks and Defenses in Federated Learning

Yangsibo Huang, Samyak Gupta, Zhao Song, Kai Li, Sanjeev Arora

Gradient inversion attack (or input recovery from gradient) is an emerging threat to the security and privacy preservation of Federated learning, whereby malicious eavesdroppers or participants in the protocol can recover (partially) the cl

ients' private data. This paper evaluates existing attacks and defenses. We find that some attacks make strong assumptions about the setup. Relaxing such assumptions can substantially weaken these attacks. We then evaluate the benefits of three proposed defense mechanisms against gradient inversion attacks. We show the trade-offs of privacy leakage and data utility of these defense methods, and find that combining them in an appropriate manner makes the attack less effective, even under the original strong assumptions. We also estimate the computation cost of end-to-end recovery of a single image under each evaluated defense. Our findings suggest that the state-of-the-art attacks can currently be defended against with minor data utility loss, as summarized in a list of potential strategies.
****
Faster Non-asymptotic Convergence for Double Q-learning
Lin Zhao, Huaqing Xiong, 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.
****
Towards Tight Communication Lower Bounds for Distributed Optimisation
Janne H. Korhonen, Dan Alistarh
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 Multi-Resolution Transformer Fine-tuning for Extreme Multi-label Text Classification
Jiong Zhang, Wei-Cheng Chang, Hsiang-Fu Yu, Inderjit Dhillon
Extreme multi-label text classification~(XMC) seeks to find relevant labels from an extreme large label collection for a given text input. Many real-world applications can be formulated as XMC problems, such as recommendation systems, document tagging and semantic search. Recently, transformer based XMC methods, such as X-Transformer and LightXML, have shown significant improvement over other XMC methods. Despite leveraging pre-trained transformer models for text representation, the fine-tuning procedure of transformer models on large label space still has lengthy computational time even with powerful GPUs. In this paper, we propose a novel recursive approach, XR-Transformer to accelerate the procedure through recursively fine-tuning transformer models on a series of multi-resolution objectives related to the original XMC objective function. Empirical results show that XR-Transformer takes significantly less training time compared to other transformer-based XMC models while yielding better state-of-the-art results. In particular, on the public Amazon-3M dataset with 3 million labels, XR-Transformer is not only 20x faster than X-Transformer but also improves the Precision@1 from 51% to 54%.
****
HRFormer: High-Resolution Vision Transformer for Dense Predict
YUHUI YUAN, Rao Fu, Lang Huang, Weihong Lin, Chao Zhang, Xilin Chen, Jingdong Wang
We present a High-Resolution Transformer (HRFormer) that learns high-resolution representations for dense prediction tasks, in contrast to the original Vision Transformer that produces low-resolution representations and has high memory and computational cost. We take advantage of the multi-resolution parallel design introduced in high-resolution convolutional networks (HRNet [45]), along with local-window self-attention that performs self-attention over small non-overlapping image windows [21], for improving the memory and computation efficiency. In addition, we introduce a convolution into the FFN to exchange information across the disconnected image windows. We demonstrate the effectiveness of the HighResolution Transformer on both human pose estimation and semantic segmentation tasks, e.g., HRFormer outperforms Swin transformer [27] by 1.3 AP on COCO pose estimatio

n with 50% fewer parameters and 30% fewer FLOPs. Code is available at: https://github.com/HRNet/HRFormer
****
Manifold Topology Divergence: a Framework for Comparing Data Manifolds.
Serguei Barannikov, Ilya Trofimov, Grigorii Sotnikov, Ekaterina Trimbach, Alexander Korotin, Alexander Filippov, Evgeny Burnaev
We propose a framework for comparing data manifolds, aimed, in particular, towards the evaluation of deep generative models. We describe a novel tool, Cross-Barcode(P,Q), that, given a pair of distributions in a high-dimensional space, tracks multiscale topology spacial discrepancies between manifolds on which the distributions are concentrated. Based on the Cross-Barcode, we introduce the Manifold Topology Divergence score (MTop-Divergence) and apply it to assess the performance of deep generative models in various domains: images, 3D-shapes, time-series, and on different datasets: MNIST, Fashion MNIST, SVHN, CIFAR10, FFHQ, market stock data, ShapeNet. We demonstrate that the MTop-Divergence accurately detects various degrees of mode-dropping, intra-mode collapse, mode invention, and image disturbance. Our algorithm scales well (essentially linearly) with the increase of the dimension of the ambient high-dimensional space. It is one of the first TDA-based methodologies that can be applied universally to datasets of different sizes and dimensions, including the ones on which the most recent GANs in the visual domain are trained. The proposed method is domain agnostic and does not rely on pre-trained networks.
****
Weak-shot Fine-grained Classification via Similarity Transfer
Junjie Chen, Li Niu, Liu Liu, Liqing Zhang
Recognizing fine-grained categories remains a challenging task, due to the subtle distinctions among different subordinate categories, which results in the need of abundant annotated samples. To alleviate the data-hungry problem, we consider the problem of learning novel categories from web data with the support of a clean set of base categories, which is referred to as weak-shot learning. In this setting, we propose a method called SimTrans to transfer pairwise semantic similarity from base categories to novel categories. Specifically, we firstly train a similarity net on clean data, and then leverage the transferred similarity to denoise web training data using two simple yet effective strategies. In addition, we apply adversarial loss on similarity net to enhance the transferability of similarity. Comprehensive experiments demonstrate the effectiveness of our weak-shot setting and our SimTrans method.
****
Shape your Space: A Gaussian Mixture Regularization Approach to Deterministic Autoencoders
Amrutha Saseendran, Kathrin Skubch, Stefan Falkner, Margret Keuper
Variational Autoencoders (VAEs) are powerful probabilistic models to learn representations of complex data distributions. One important limitation of VAEs is the strong prior assumption that latent representations learned by the model follow a simple uni-modal Gaussian distribution. Further, the variational training procedure poses considerable practical challenges. Recently proposed regularized autoencoders offer a deterministic autoencoding framework, that simplifies the original VAE objective and is significantly easier to train. Since these models only provide weak control over the learned latent distribution, they require an ex-post density estimation step to generate samples comparable to those of VAEs. In this paper, we propose a simple and end-to-end trainable deterministic autoencoding framework, that efficiently shapes the latent space of the model during training and utilizes the capacity of expressive multi-modal latent distributions. The proposed training procedure provides direct evidence if the latent distribution adequately captures complex aspects of the encoded data. We show in experiments the expressiveness and sample quality of our model in various challenging continuous and discrete domains. An implementation is available at https://github.com/boschresearch/GMM_DAE.
****
An Even More Optimal Stochastic Optimization Algorithm: Minibatching and Interpo

lation Learning
Blake E. Woodworth, Nathan Srebro
We present and analyze an algorithm for optimizing smooth and convex or strongly convex objectives using minibatch stochastic gradient estimates. The algorithm is optimal with respect to its dependence on both the minibatch size and minimum expected loss simultaneously. This improves over the optimal method of Lan, which is insensitive to the minimum expected loss; over the optimistic acceleration of Cotter et al., which has suboptimal dependence on the minibatch size; and over the algorithm of Liu and Belkin, which is limited to least squares problems and is also similarly suboptimal. Applied to interpolation learning, the improvement over Cotter et al.~and Liu and Belkin translates to a linear, rather than square-root, parallelization speedup.
****
Indexed Minimum Empirical Divergence for Unimodal Bandits
Hassan SABER, Pierre Ménard, Odalric-Ambrym Maillard
We consider a stochastic multi-armed bandit problem specified by a set of one-dimensional family exponential distributions endowed with a unimodal structure. The unimodal structure is of practical relevance for several applications. We introduce IMED-UB, an algorithm that exploits provably optimally the unimodal-structure, by adapting to this setting the Indexed Minimum Empirical Divergence (IMED) algorithm introduced by Honda and Takemura (2015). Owing to our proof technique, we are able to provide a concise finite-time analysis of the IMED-UB algorithm, that is simple and yet yields asymptotic optimality. We finally provide numerical experiments showing that IMED-UB competes favorably with the recently introduced state-of-the-art algorithms.
****
SOAT: A Scene- and Object-Aware Transformer for Vision-and-Language Navigation
Abhinav Moudgil, Arjun Majumdar, Harsh Agrawal, Stefan Lee, Dhruv Batra
Natural language instructions for visual navigation often use scene descriptions (e.g., bedroom) and object references (e.g., green chairs) to provide a breadcrumb trail to a goal location. This work presents a transformer-based vision-and-language navigation (VLN) agent that uses two different visual encoders -- a scene classification network and an object detector -- which produce features that match these two distinct types of visual cues. In our method, scene features contribute high-level contextual information that supports object-level processing. With this design, our model is able to use vision-and-language pretraining (i.e., learning the alignment between images and text from large-scale web data) to substantially improve performance on the Room-to-Room (R2R) and Room-Across-Room (RxR) benchmarks. Specifically, our approach leads to improvements of 1.8% absolute in SPL on R2R and 3.7% absolute in SR on RxR. Our analysis reveals even larger gains for navigation instructions that contain six or more object references, which further suggests that our approach is better able to use object features and align them to references in the instructions.
****
A Normative and Biologically Plausible Algorithm for Independent Component Analysis
Yanis Bahroun, Dmitri Chklovskii, Anirvan Sengupta
The brain effortlessly solves blind source separation (BSS) problems, but the algorithm it uses remains elusive. In signal processing, linear BSS problems are often solved by Independent Component Analysis (ICA). To serve as a model of a biological circuit, the ICA neural network (NN) must satisfy at least the following requirements: 1. The algorithm must operate in the online setting where data samples are streamed one at a time, and the NN computes the sources on the fly without storing any significant fraction of the data in memory. 2. The synaptic weight update is local, i.e., it depends only on the biophysical variables present in the vicinity of a synapse. Here, we propose a novel objective function for ICA from which we derive a biologically plausible NN, including both the neural architecture and the synaptic learning rules. Interestingly, our algorithm relies on modulating synaptic plasticity by the total activity of the output neurons. In the brain, this could be accomplished by neuromodulators, extracellular calci

um, local field potential, or nitric oxide.
****
Regret Bounds for Gaussian-Process Optimization in Large Domains
Manuel Wuethrich, Bernhard Schölkopf, Andreas Krause
The goal of this paper is to characterize Gaussian-Process optimization in the setting where the function domain is large relative to the number of admissible function evaluations, i.e., where it is impossible to find the global optimum. We provide upper bounds on the suboptimality (Bayesian simple regret) of the solution found by optimization strategies that are closely related to the widely used expected improvement (EI) and upper confidence bound (UCB) algorithms. These regret bounds illuminate the relationship between the number of evaluations, the domain size (i.e. cardinality of finite domains / Lipschitz constant of the covariance function in continuous domains), and the optimality of the retrieved function value.In particular, we show that even when the number of evaluations is far too small to find the global optimum, we can find nontrivial function values (e.g. values that achieve a certain ratio with the optimal value).
****
Deeply Shared Filter Bases for Parameter-Efficient Convolutional Neural Networks
Woochul Kang, Daeyeon Kim
Modern convolutional neural networks (CNNs) have massive identical convolution blocks, and, hence, recursive sharing of parameters across these blocks has been proposed to reduce the amount of parameters. However, naive sharing of parameters poses many challenges such as limited representational power and the vanishing/exploding gradients problem of recursively shared parameters. In this paper, we present a recursive convolution block design and training method, in which a recursively shareable part, or a filter basis, is separated and learned while effectively avoiding the vanishing/exploding gradients problem during training. We show that the unwieldy vanishing/exploding gradients problem can be controlled by enforcing the elements of the filter basis orthonormal, and empirically demonstrate that the proposed orthogonality regularization improves the flow of gradients during training. Experimental results on image classification and object detection show that our approach, unlike previous parameter-sharing approaches, does not trade performance to save parameters and consistently outperforms over parameterized counterpart networks. This superior performance demonstrates that the proposed recursive convolution block design and the orthogonality regularization not only prevent performance degradation, but also consistently improve the representation capability while a significant amount of parameters are recursively shared.
****
On Optimal Robustness to Adversarial Corruption in Online Decision Problems
Shinji Ito
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Directed Spectrum Measures Improve Latent Network Models Of Neural Populations
Neil Gallagher, Kafui Dzirasa, David Carlson
Systems neuroscience aims to understand how networks of neurons distributed throughout the brain mediate computational tasks. One popular approach to identify those networks is to first calculate measures of neural activity (e.g. power spectra) from multiple brain regions, and then apply a linear factor model to those measures. Critically, despite the established role of directed communication between brain regions in neural computation, measures of directed communication have been rarely utilized in network estimation because they are incompatible with the implicit assumptions of the linear factor model approach. Here, we develop a novel spectral measure of directed communication called the Directed Spectrum (DS). We prove that it is compatible with the implicit assumptions of linear factor models, and we provide a method to estimate the DS. We demonstrate that laten

t linear factor models of DS measures better capture underlying brain networks i
n both simulated and real neural recording data compared to available alternativ
es. Thus, linear factor models of the Directed Spectrum offer neuroscientists a
simple and effective way to explicitly model directed communication in networks
of neural populations.
****
Uncertainty-Based Offline Reinforcement Learning with Diversified Q-Ensemble
Gaon An, Seungyong Moon, Jang-Hyun Kim, Hyun Oh Song
Offline reinforcement learning (offline RL), which aims to find an optimal polic
y from a previously collected static dataset, bears algorithmic difficulties due
 to function approximation errors from out-of-distribution (OOD) data points. To
 this end, offline RL algorithms adopt either a constraint or a penalty term tha
t explicitly guides the policy to stay close to the given dataset. However, prio
r methods typically require accurate estimation of the behavior policy or sampli
ng from OOD data points, which themselves can be a non-trivial problem. Moreover
, these methods under-utilize the generalization ability of deep neural networks
 and often fall into suboptimal solutions too close to the given dataset. In thi
s work, we propose an uncertainty-based offline RL method that takes into accoun
t the confidence of the Q-value prediction and does not require any estimation o
r sampling of the data distribution. We show that the clipped Q-learning, a tech
nique widely used in online RL, can be leveraged to successfully penalize OOD da
ta points with high prediction uncertainties. Surprisingly, we find that it is p
ossible to substantially outperform existing offline RL methods on various tasks
 by simply increasing the number of Q-networks along with the clipped Q-learning
. Based on this observation, we propose an ensemble-diversified actor-critic alg
orithm that reduces the number of required ensemble networks down to a tenth com
pared to the naive ensemble while achieving state-of-the-art performance on most
 of the D4RL benchmarks considered.
****
Distribution-free inference for regression: discrete, continuous, and in between
Yonghoon Lee, Rina Barber
Requests for name changes in the electronic proceedings will be accepted with 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.
****
Statistical Inference with M-Estimators on Adaptively Collected Data
Kelly Zhang, Lucas Janson, Susan Murphy
Bandit algorithms are increasingly used in real-world sequential decision-making
 problems. Associated with this is an increased desire to be able to use the res
ulting datasets to answer scientific questions like: Did one type of ad lead to
more purchases? In which contexts is a mobile health intervention effective? How
ever, classical statistical approaches fail to provide valid confidence interval
s when used with data collected with bandit algorithms. Alternative methods have
 recently been developed for simple models (e.g., comparison of means). Yet ther
e is a lack of general methods for conducting statistical inference using more
complex models on data collected with (contextual) bandit algorithms; for exampl
e, current methods cannot be used for valid inference on parameters in a logisti
c regression model for a binary reward. In this work, we develop theory justify
ing the use of M-estimators---which includes estimators based on empirical risk
 minimization as well as maximum likelihood---on data collected with adaptive al
gorithms, including (contextual) bandit algorithms. Specifically, we show that M
-estimators, modified with particular adaptive weights, can be used to construc
t asymptotically valid confidence regions for a variety of inferential targets.
****
NeuroLKH: Combining Deep Learning Model with Lin-Kernighan-Helsgaun Heuristic fo
r Solving the Traveling Salesman Problem
Liang Xin, Wen Song, Zhiguang Cao, Jie Zhang
We present NeuroLKH, a novel algorithm that combines deep learning with the stro
ng traditional heuristic Lin-Kernighan-Helsgaun (LKH) for solving Traveling Sale

sman Problem. Specifically, we train a Sparse Graph Network (SGN) with supervise
d learning for edge scores and unsupervised learning for node penalties, both of
 which are critical for improving the performance of LKH. Based on the output of
 SGN, NeuroLKH creates the edge candidate set and transforms edge distances to g
uide the searching process of LKH. Extensive experiments firmly demonstrate that
, by training one model on a wide range of problem sizes, NeuroLKH significantly
 outperforms LKH and generalizes well to much larger sizes. Also, we show that N
euroLKH can be applied to other routing problems such as Capacitated Vehicle Rou
ting Problem (CVRP), Pickup and Delivery Problem (PDP), and CVRP with Time Windo
ws (CVRPTW).
****
LSH-SMILE: Locality Sensitive Hashing Accelerated Simulation and Learning
Chonghao Sima, Yexiang Xue
The advancement of deep neural networks over the last decade has enabled progres
s in scientific knowledge discovery in the form of learning Partial Differential
 Equations (PDEs) directly from experiment data. Nevertheless, forward simulatio
n and backward learning of large-scale dynamic systems require handling billions
 of mutually interacting elements, the scale of which overwhelms current computi
ng architectures. We propose Locality Sensitive Hashing Accelerated Simulation a
nd Learning (LSH-SMILE), a unified framework to scale up both forward simulation
 and backward learning of physics systems. LSH-SMILE takes advantage of (i) the
locality of PDE updates, (ii) similar temporal dynamics shared by multiple eleme
nts. LSH-SMILE hashes elements with similar dynamics into a single hash bucket a
nd handles their updates at once. This allows LSH-SMILE to scale with respect to
 the number of non-empty hash buckets, a drastic improvement over conventional a
pproaches. Theoretically, we prove a novel bound on the errors introduced by LSH
-SMILE. Experimentally, we demonstrate that LSH-SMILE simulates physics systems
at comparable quality with exact approaches, but with way less time and space co
mplexity. Such savings also translate to better learning performance due to LSH-
SMILE's ability to propagate gradients over a long duration.
****
Meta-learning with an Adaptive Task Scheduler
Huaxiu Yao, Yu Wang, Ying Wei, Peilin Zhao, Mehrdad Mahdavi, Defu Lian, Chelsea
Finn
To benefit the learning of a new task, meta-learning has been proposed to transf
er a well-generalized meta-model learned from various meta-training tasks. Exist
ing meta-learning algorithms randomly sample meta-training tasks with a uniform
probability, under the assumption that tasks are of equal importance. However, i
t is likely that tasks are detrimental with noise or imbalanced given a limited
number of meta-training tasks. To prevent the meta-model from being corrupted by
 such detrimental tasks or dominated by tasks in the majority, in this paper, we
 propose an adaptive task scheduler (ATS) for the meta-training process. In ATS,
 for the first time, we design a neural scheduler to decide which meta-training
tasks to use next by predicting the probability being sampled for each candidate
 task, and train the scheduler to optimize the generalization capacity of the me
ta-model to unseen tasks. We identify two meta-model-related factors as the inpu
t of the neural scheduler, which characterize the difficulty of a candidate task
 to the meta-model. Theoretically, we show that a scheduler taking the two facto
rs into account improves the meta-training loss and also the optimization landsc
ape. Under the setting of meta-learning with noise and limited budgets, ATS impr
oves the performance on both miniImageNet and a real-world drug discovery benchm
ark by up to 13% and 18%, respectively, compared to state-of-the-art task schedu
lers.
****
Neural Active Learning with Performance Guarantees
Zhilei Wang, Pranjal Awasthi, Christoph Dann, Ayush Sekhari, Claudio Gentile
We investigate the problem of active learning in the streaming setting in non-pa
rametric regimes, where the labels are stochastically generated from a class of
functions on which we make no assumptions whatsoever. We rely on recently propos
ed Neural Tangent Kernel (NTK) approximation tools to construct a suitable neura

l embedding that determines the feature space the algorithm operates on and the learned model computed atop. Since the shape of the label requesting threshold is tightly related to the complexity of the function to be learned, which is a-priori unknown, we also derive a version of the algorithm which is agnostic to any prior knowledge. This algorithm relies on a regret balancing scheme to solve the resulting online model selection problem, and is computationally efficient. We prove joint guarantees on the cumulative regret and number of requested labels which depend on the complexity of the labeling function at hand. In the linear case, these guarantees recover known minimax results of the generalization error as a function of the label complexity in a standard statistical learning setting.
****
A Gradient Method for Multilevel Optimization
Ryo Sato, Mirai Tanaka, Akiko Takeda
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Edge Representation Learning with Hypergraphs
Jaehyeong Jo, Jinheon Baek, Seul Lee, Dongki Kim, Minki Kang, Sung Ju Hwang
Graph neural networks have recently achieved remarkable success in representing graph-structured data, with rapid progress in both the node embedding and graph pooling methods. Yet, they mostly focus on capturing information from the nodes considering their connectivity, and not much work has been done in representing the edges, which are essential components of a graph. However, for tasks such as graph reconstruction and generation, as well as graph classification tasks for which the edges are important for discrimination, accurately representing edges of a given graph is crucial to the success of the graph representation learning. To this end, we propose a novel edge representation learning framework based on Dual Hypergraph Transformation (DHT), which transforms the edges of a graph into the nodes of a hypergraph. This dual hypergraph construction allows us to apply message-passing techniques for node representations to edges. After obtaining edge representations from the hypergraphs, we then cluster or drop edges to obtain holistic graph-level edge representations. We validate our edge representation learning method with hypergraphs on diverse graph datasets for graph representation and generation performance, on which our method largely outperforms existing graph representation learning methods. Moreover, our edge representation learning and pooling method also largely outperforms state-of-the-art graph pooling methods on graph classification, not only because of its accurate edge representation learning, but also due to its lossless compression of the nodes and removal of irrelevant edges for effective message-passing.
****
One Question Answering Model for Many Languages with Cross-lingual Dense Passage Retrieval
Akari Asai, Xinyan Yu, Jungo Kasai, Hanna Hajishirzi
We present Cross-lingual Open-Retrieval Answer Generation (CORA), the first unified many-to-many question answering (QA) model that can answer questions across many languages, even for ones without language-specific annotated data or knowledge sources.We introduce a new dense passage retrieval algorithm that is trained to retrieve documents across languages for a question.Combined with a multilingual autoregressive generation model, CORA answers directly in the target language without any translation or in-language retrieval modules as used in prior work. We propose an iterative training method that automatically extends annotated data available only in high-resource languages to low-resource ones. Our results show that CORA substantially outperforms the previous state of the art on multilingual open QA benchmarks across 26 languages, 9 of which are unseen during training. Our analyses show the significance of cross-lingual retrieval and generation in many languages, particularly under low-resource settings.
****

LEADS: Learning Dynamical Systems that Generalize Across Environments
Yuan Yin, Ibrahim Ayed, Emmanuel de Bézenac, Nicolas Baskiotis, Patrick Gallinar
i
When modeling dynamical systems from real-world data samples, the distribution o
f data often changes according to the environment in which they are captured, an
d the dynamics of the system itself vary from one environment to another. Genera
lizing across environments thus challenges the conventional frameworks. The clas
sical settings suggest either considering data as i.i.d and learning a single mo
del to cover all situations or learning environment-specific models. Both are su
b-optimal: the former disregards the discrepancies between environments leading
to biased solutions, while the latter does not exploit their potential commonali
ties and is prone to scarcity problems. We propose LEADS, a novel framework that
 leverages the commonalities and discrepancies among known environments to impro
ve model generalization. This is achieved with a tailored training formulation a
iming at capturing common dynamics within a shared model while additional terms
capture environment-specific dynamics. We ground our approach in theory, exhibit
ing a decrease in sample complexity w.r.t classical alternatives. We show how t
heory and practice coincides on the simplified case of linear dynamics. Moreover
, we instantiate this framework for neural networks and evaluate it experimental
ly on representative families of nonlinear dynamics. We show that this new setti
ng can exploit knowledge extracted from environment-dependent data and improves
generalization for both known and novel environments.
****
Storchastic: A Framework for General Stochastic Automatic Differentiation
Emile Krieken, Jakub Tomczak, Annette Ten Teije
Modelers use automatic differentiation (AD) of computation graphs to implement c
omplex Deep Learning models without defining gradient computations. Stochastic A
D extends AD to stochastic computation graphs with sampling steps, which arise w
hen modelers handle the intractable expectations common in Reinforcement Learnin
g and Variational Inference. However, current methods for stochastic AD are limi
ted: They are either only applicable to continuous random variables and differen
tiable functions, or can only use simple but high variance score-function estima
tors. To overcome these limitations, we introduce Storchastic, a new framework f
or AD of stochastic computation graphs. Storchastic allows the modeler to choose
 from a wide variety of gradient estimation methods at each sampling step, to op
timally reduce the variance of the gradient estimates. Furthermore, Storchastic
is provably unbiased for estimation of any-order gradients, and generalizes vari
ance reduction techniques to higher-order gradient estimates. Finally, we implem
ent Storchastic as a PyTorch library at github.com/HEmile/storchastic.
****
Concentration inequalities under sub-Gaussian and sub-exponential conditions
Andreas Maurer, Massimiliano Pontil
We prove analogues of the popular bounded difference inequality (also called McD
iarmid's inequality) for functions of independent random variables under sub-gau
ssian and sub-exponential conditions. Applied to vector-valued concentration and
 the method of Rademacher complexities these inequalities allow an easy extensio
n of uniform convergence results for PCA and linear regression to the case poten
tially unbounded input- and output variables.
****
Variance-Aware Off-Policy Evaluation with Linear Function Approximation
Yifei Min, Tianhao Wang, Dongruo Zhou, Quanquan Gu
We study the off-policy evaluation (OPE) problem in reinforcement learning with
linear function approximation, which aims to estimate the value function of a ta
rget policy based on the offline data collected by a behavior policy. We propose
 to incorporate the variance information of the value function to improve the sa
mple efficiency of OPE. More specifically, for time-inhomogeneous episodic linea
r Markov decision processes (MDPs), we propose an algorithm, \texttt{VA-OPE}, wh
ich uses the estimated variance of the value function to reweight the Bellman re
sidual in Fitted Q-Iteration. We show that our algorithm achieves a tighter erro
r bound than the best-known result. We also provide a fine-grained characterizat

ion of the distribution shift between the behavior policy and the target policy.
Extensive numerical experiments corroborate our theory.
****
A Provably Efficient Sample Collection Strategy for Reinforcement Learning
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.
****
Improved Regret Bounds for Tracking Experts with Memory
James Robinson, Mark Herbster
We address the problem of sequential prediction with expert advice in a non-stat
ionary environment with long-term memory guarantees in the sense of Bousquet and
Warmuth [4]. We give a linear-time algorithm that improves on the best known re
gret bound [27]. This algorithm incorporates a relative entropy projection step.
This projection is advantageous over previous weight-sharing approaches in that
weight updates may come with implicit costs as in for example portfolio optimiz
ation. We give an algorithm to compute this projection step in linear time, whic
h may be of independent interest.
****
Robustness of Graph Neural Networks at Scale
Simon Geisler, Tobias Schmidt, Hakan ■irin, Daniel Zügner, Aleksandar Bojchevski
, Stephan Günnemann
Graph Neural Networks (GNNs) are increasingly important given their popularity a
nd the diversity of applications. Yet, existing studies of their vulnerability t
o adversarial attacks rely on relatively small graphs. We address this gap and s
tudy how to attack and defend GNNs at scale. We propose two sparsity-aware first
-order optimization attacks that maintain an efficient representation despite op
timizing over a number of parameters which is quadratic in the number of nodes.
We show that common surrogate losses are not well-suited for global attacks on G
NNs. Our alternatives can double the attack strength. Moreover, to improve GNNs'
reliability we design a robust aggregation function, Soft Median, resulting in
an effective defense at all scales. We evaluate our attacks and defense with sta
ndard GNNs on graphs more than 100 times larger compared to previous work. We ev
en scale one order of magnitude further by extending our techniques to a scalabl
e GNN.
****
Random Noise Defense Against Query-Based Black-Box Attacks
Zeyu Qin, Yanbo Fan, Hongyuan Zha, Baoyuan Wu
The query-based black-box attacks have raised serious threats to machine learnin
g models in many real applications. In this work, we study a lightweight defense
method, dubbed Random Noise Defense (RND), which adds proper Gaussian noise to
each query. We conduct the theoretical analysis about the effectiveness of RND a
gainst query-based black-box attacks and the corresponding adaptive attacks. Our
theoretical results reveal that the defense performance of RND is determined by
the magnitude ratio between the noise induced by RND and the noise added by the
attackers for gradient estimation or local search. The large magnitude ratio l
eads to the stronger defense performance of RND, and it's also critical for miti
gating adaptive attacks. Based on our analysis, we further propose to combine RN
D with a plausible Gaussian augmentation Fine-tuning (RND-GF). It enables RND to
add larger noise to each query while maintaining the clean accuracy to obtain a
better trade-off between clean accuracy and defense performance. Additionally,
RND can be flexibly combined with the existing defense methods to further boost
the adversarial robustness, such as adversarial training (AT). Extensive experim
ents on CIFAR-10 and ImageNet verify our theoretical findings and the effectiven
ess of RND and RND-GF.
****
SADGA: Structure-Aware Dual Graph Aggregation Network for Text-to-SQL
Ruichu Cai, Jinjie Yuan, Boyan Xu, Zhifeng Hao

The Text-to-SQL task, aiming to translate the natural language of the questions into SQL queries, has drawn much attention recently. One of the most challengin g problems of Text-to-SQL is how to generalize the trained model to the unseen d atabase schemas, also known as the cross-domain Text-to-SQL task. The key lies i n the generalizability of (i) the encoding method to model the question and the database schema and (ii) the question-schema linking method to learn the mapping between words in the question and tables/columns in the database schema. Focusi ng on the above two key issues, we propose a \emph{Structure-Aware Dual Graph Ag gregation Network} (SADGA) for cross-domain Text-to-SQL. In SADGA, we adopt the graph structure to provide a unified encoding model for both the natural languag e question and database schema. Based on the proposed unified modeling, we furth er devise a structure-aware aggregation method to learn the mapping between the question-graph and schema-graph. The structure-aware aggregation method is featu red with \emph{Global Graph Linking}, \emph{Local Graph Linking} and \emph{Dual-Graph Aggregation Mechanism}. We not only study the performance of our proposal empirically but also achieved 3rd place on the challenging Text-to-SQL benchmark Spider at the time of writing.
****
Near-Optimal Offline Reinforcement Learning via Double Variance Reduction
Ming Yin, Yu Bai, 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.
****
Joint Modeling of Visual Objects and Relations for Scene Graph Generation
Minghao Xu, Meng Qu, Bingbing Ni, Jian Tang
An in-depth scene understanding usually requires recognizing all the objects and their relations in an image, encoded as a scene graph. Most existing approaches for scene graph generation first independently recognize each object and then p redict their relations independently. Though these approaches are very efficient, they ignore the dependency between different objects as well as between their relations. In this paper, we propose a principled approach to jointly predict th e entire scene graph by fully capturing the dependency between different objects and between their relations. Specifically, we establish a unified conditional r andom field (CRF) to model the joint distribution of all the objects and their r elations in a scene graph. We carefully design the potential functions to enable relational reasoning among different objects according to knowledge graph embed ding methods. We further propose an efficient and effective algorithm for infere nce based on mean-field variational inference, in which we first provide a warm initialization by independently predicting the objects and their relations accor ding to the current model, followed by a few iterations of relational reasoning. Experimental results on both the relationship retrieval and zero-shot relations hip retrieval tasks prove the efficiency and efficacy of our proposed approach.
****
Going Beyond Linear Transformers with Recurrent Fast Weight Programmers
Kazuki Irie, Imanol Schlag, Róbert Csordás, Jürgen Schmidhuber
Transformers with linearised attention (''linear Transformers'') have demonstrat ed the practical scalability and effectiveness of outer product-based Fast Weigh t Programmers (FWPs) from the '90s. However, the original FWP formulation is mor e general than the one of linear Transformers: a slow neural network (NN) contin ually reprograms the weights of a fast NN with arbitrary architecture. In existi ng linear Transformers, both NNs are feedforward and consist of a single layer. Here we explore new variations by adding recurrence to the slow and fast nets. W e evaluate our novel recurrent FWPs (RFWPs) on two synthetic algorithmic tasks (code execution and sequential ListOps), Wikitext-103 language models, and on the Atari 2600 2D game environment. Our models exhibit properties of Transformers a nd RNNs. In the reinforcement learning setting, we report large improvements ove r LSTM in several Atari games. Our code is public.
****

Reinforced Few-Shot Acquisition Function Learning for Bayesian Optimization
Bing-Jing Hsieh, Ping-Chun Hsieh, Xi Liu
Bayesian optimization (BO) conventionally relies on handcrafted acquisition functions (AFs) to sequentially determine the sample points. However, it has been widely observed in practice that the best-performing AF in terms of regret can vary significantly under different types of black-box functions. It has remained a challenge to design one AF that can attain the best performance over a wide variety of black-box functions. This paper aims to attack this challenge through the perspective of reinforced few-shot AF learning (FSAF). Specifically, we first connect the notion of AFs with Q-functions and view a deep Q-network (DQN) as a surrogate differentiable AF. While it serves as a natural idea to combine DQN and an existing few-shot learning method, we identify that such a direct combination does not perform well due to severe overfitting, which is particularly critical in BO due to the need of a versatile sampling policy. To address this, we present a Bayesian variant of DQN with the following three features: (i) It learns a distribution of Q-networks as AFs based on the Kullback-Leibler regularization framework. This inherently provides the uncertainty required in sampling for BO and mitigates overfitting. (ii) For the prior of the Bayesian DQN, we propose to use a demo policy induced by an off-the-shelf AF for better training stability. (iii) On the meta-level, we leverage the meta-loss of Bayesian model-agnostic meta-learning, which serves as a natural companion to the proposed FSAF. Moreover, with the proper design of the Q-networks, FSAF is general-purpose in that it is agnostic to the dimension and the cardinality of the input domain. Through extensive experiments, we demonstrate that the FSAF achieves comparable or better regrets than the state-of-the-art benchmarks on a wide variety of synthetic and real-world test functions.
****
Forster Decomposition and Learning Halfspaces with Noise
Ilias Diakonikolas, Daniel Kane, Christos Tzamos
A Forster transform is an operation that turns a multivariate distribution into one with good anti-concentration properties. While a Forster transform does not always exist, we show that any distribution can be efficiently decomposed as a disjoint mixture of few distributions for which a Forster transform exists and can be computed efficiently. As the main application of this result, we obtain the first polynomial-time algorithm for distribution-independent PAC learning of halfspaces in the Massart noise model with strongly polynomial sample complexity, i.e., independent of the bit complexity of the examples. Previous algorithms for this learning problem incurred sample complexity scaling polynomially with the bit complexity, even though such a dependence is not information-theoretically necessary.
****
Cortico-cerebellar networks as decoupling neural interfaces
Joseph Pemberton, Ellen Boven, Richard Apps, Rui Ponte Costa
The brain solves the credit assignment problem remarkably well. For credit to be assigned across neural networks they must, in principle, wait for specific neural computations to finish. How the brain deals with this inherent locking problem has remained unclear. Deep learning methods suffer from similar locking constraints both on the forward and feedback phase. Recently, decoupled neural interfaces (DNIs) were introduced as a solution to the forward and feedback locking problems in deep networks.Here we propose that a specialised brain region, the cerebellum, helps the cerebral cortex solve similar locking problems akin to DNIs. To demonstrate the potential of this framework we introduce a systems-level model in which a recurrent cortical network receives online temporal feedback predictions from a cerebellar module. We test this cortico-cerebellar recurrent neural network (ccRNN) model on a number of sensorimotor (line and digit drawing) and cognitive tasks (pattern recognition and caption generation) that have been shown to be cerebellar-dependent. In all tasks, we observe that ccRNNs facilitates learning while reducing ataxia-like behaviours, consistent with classical experimental observations. Moreover, our model also explains recent behavioural and neuronal observations while making several testable predictions across multiple leve

ls.Overall, our work offers a novel perspective on the cerebellum as a brain-wide decoupling machine for efficient credit assignment and opens a new avenue between deep learning and neuroscience.
****

To The Point: Correspondence-driven monocular 3D category reconstruction
Filippos Kokkinos, Iasonas Kokkinos
We present To The Point (TTP), a method for reconstructing 3D objects from a single image using 2D to 3D correspondences given only foreground masks, a category specific template and optionally sparse keypoints for supervision. We recover a 3D shape from a 2D image by first regressing the 2D positions corresponding to the 3D template vertices and then jointly estimating a rigid camera transform and non-rigid template deformation that optimally explain the 2D positions through the 3D shape projection. By relying on correspondences we use a simple per-sample optimization problem to replace CNN-based regression of camera pose and non-rigid deformation and thereby obtain substantially more accurate 3D reconstructions. We treat this optimization as a differentiable layer and train the whole system in an end-to-end manner using geometry-driven losses. We report systematic quantitative improvements on multiple categories and provide qualitative results comprising diverse shape, poses and texture prediction examples.
****

Proper Value Equivalence
Christopher Grimm, Andre Barreto, Greg Farquhar, David Silver, Satinder Singh
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

Challenges and Opportunities in High Dimensional Variational Inference
Akash Kumar Dhaka, Alejandro Catalina, Manushi Welandawe, Michael R. Andersen, Jonathan Huggins, Aki Vehtari
Current black-box variational inference (BBVI) methods require the user to make numerous design choices – such as the selection of variational objective and approximating family – yet there is little principled guidance on how to do so. We develop a conceptual framework and set of experimental tools to understand the effects of these choices, which we leverage to propose best practices for maximizing posterior approximation accuracy. Our approach is based on studying the pre-asymptotic tail behavior of the density ratios between the joint distribution and the variational approximation, then exploiting insights and tools from the importance sampling literature. Our framework and supporting experiments help to distinguish between the behavior of BBVI methods for approximating low-dimensional versus moderate-to-high-dimensional posteriors. In the latter case, we show that mass-covering variational objectives are difficult to optimize and do not improve accuracy, but flexible variational families can improve accuracy and the effectiveness of importance sampling – at the cost of additional optimization challenges. Therefore, for moderate-to-high-dimensional posteriors we recommend using the (mode-seeking) exclusive KL divergence since it is the easiest to optimize, and improving the variational family or using model parameter transformations to make the posterior and optimal variational approximation more similar. On the other hand, in low-dimensional settings, we show that heavy-tailed variational families and mass-covering divergences are effective and can increase the chances that the approximation can be improved by importance sampling.
****

On the Expressivity of Markov Reward
David Abel, Will Dabney, Anna Harutyunyan, Mark K. Ho, Michael Littman, Doina Precup, Satinder Singh
Reward is the driving force for reinforcement-learning agents. This paper is dedicated to understanding the expressivity of reward as a way to capture tasks that we would want an agent to perform. We frame this study around three new abstract notions of “task” that might be desirable: (1) a set of acceptable behaviors, (2) a partial ordering over behaviors, or (3) a partial ordering over trajector

ies. Our main results prove that while reward can express many of these tasks, there exist instances of each task type that no Markov reward function can capture. We then provide a set of polynomial-time algorithms that construct a Markov reward function that allows an agent to optimize tasks of each of these three types, and correctly determine when no such reward function exists. We conclude with an empirical study that corroborates and illustrates our theoretical findings.
****

One More Step Towards Reality: Cooperative Bandits with Imperfect Communication
Udari Madhushani, Abhimanyu Dubey, Naomi Leonard, Alex Pentland
The cooperative bandit problem is increasingly becoming relevant due to its applications in large-scale decision-making. However, most research for this problem focuses exclusively on the setting with perfect communication, whereas in most real-world distributed settings, communication is often over stochastic networks, with arbitrary corruptions and delays. In this paper, we study cooperative bandit learning under three typical real-world communication scenarios, namely, (a) message-passing over stochastic time-varying networks, (b) instantaneous reward-sharing over a network with random delays, and (c) message-passing with adversarially corrupted rewards, including byzantine communication. For each of these environments, we propose decentralized algorithms that achieve competitive performance, along with near-optimal guarantees on the incurred group regret as well. Furthermore, in the setting with perfect communication, we present an improved delayed-update algorithm that outperforms the existing state-of-the-art on various network topologies. Finally, we present tight network-dependent minimax lower bounds on the group regret. Our proposed algorithms are straightforward to implement and obtain competitive empirical performance.
****

Multi-Agent Reinforcement Learning in Stochastic Networked Systems
Yiheng Lin, Guannan Qu, Longbo Huang, Adam Wierman
We study multi-agent reinforcement learning (MARL) in a stochastic network of agents. The objective is to find localized policies that maximize the (discounted) global reward. In general, scalability is a challenge in this setting because the size of the global state/action space can be exponential in the number of agents. Scalable algorithms are only known in cases where dependencies are static, fixed and local, e.g., between neighbors in a fixed, time-invariant underlying graph. In this work, we propose a Scalable Actor Critic framework that applies in settings where the dependencies can be non-local and stochastic, and provide a finite-time error bound that shows how the convergence rate depends on the speed of information spread in the network. Additionally, as a byproduct of our analysis, we obtain novel finite-time convergence results for a general stochastic approximation scheme and for temporal difference learning with state aggregation, which apply beyond the setting of MARL in networked systems.
****

Neural Scene Flow Prior
Xueqian Li, Jhony Kaesemodel Pontes, Simon Lucey
Before the deep learning revolution, many perception algorithms were based on runtime optimization in conjunction with a strong prior/regularization penalty. A prime example of this in computer vision is optical and scene flow. Supervised learning has largely displaced the need for explicit regularization. Instead, they rely on large amounts of labeled data to capture prior statistics, which are not always readily available for many problems. Although optimization is employed to learn the neural network, at runtime, the weights of this network are frozen. As a result, these learning solutions are domain-specific and do not generalize well to other statistically different scenarios. This paper revisits the scene flow problem that relies predominantly on runtime optimization and strong regularization. A central innovation here is the inclusion of a neural scene flow prior, which utilizes the architecture of neural networks as a new type of implicit regularizer. Unlike learning-based scene flow methods, optimization occurs at runtime, and our approach needs no offline datasets---making it ideal for deployment in new environments such as autonomous driving. We show that an architecture based exclusively on multilayer perceptrons (MLPs) can be used as a scene flow

prior. Our method attains competitive---if not better---results on scene flow b
enchmarks. Also, our neural prior's implicit and continuous scene flow represent
ation allows us to estimate dense long-term correspondences across a sequence of
 point clouds. The dense motion information is represented by scene flow fields
where points can be propagated through time by integrating motion vectors. We de
monstrate such a capability by accumulating a sequence of lidar point clouds.
****
The future is log-Gaussian: ResNets and their infinite-depth-and-width limit at
initialization
Mufan Li, Mihai Nica, Dan Roy
Requests for name changes in the electronic proceedings will be accepted with 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.
****
Grammar-Based Grounded Lexicon Learning
Jiayuan Mao, Freda Shi, Jiajun Wu, Roger Levy, Josh Tenenbaum
Requests for name changes in the electronic proceedings will be accepted with 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.
****
Distributed Deep Learning In Open Collaborations
Michael Diskin, Alexey Bukhtiyarov, Max Ryabinin, Lucile Saulnier, quentin lhoes
t, Anton Sinitsin, Dmitry Popov, Dmitry V. Pyrkin, Maxim Kashirin, Alexander Bor
zunov, Albert Villanova del Moral, Denis Mazur, Ilia Kobelev, Yacine Jernite, Th
omas Wolf, Gennady Pekhimenko
Modern deep learning applications require increasingly more compute to train sta
te-of-the-art models. To address this demand, large corporations and institution
s use dedicated High-Performance Computing clusters, whose construction and main
tenance are both environmentally costly and well beyond the budget of most organ
izations. As a result, some research directions become the exclusive domain of a
 few large industrial and even fewer academic actors. To alleviate this disparit
y, smaller groups may pool their computational resources and run collaborative e
xperiments that benefit all participants. This paradigm, known as grid- or volun
teer computing, has seen successful applications in numerous scientific areas. H
owever, using this approach for machine learning is difficult due to high latenc
y, asymmetric bandwidth, and several challenges unique to volunteer computing. I
n this work, we carefully analyze these constraints and propose a novel algorith
mic framework designed specifically for collaborative training. We demonstrate t
he effectiveness of our approach for SwAV and ALBERT pretraining in realistic co
nditions and achieve performance comparable to traditional setups at a fraction
of the cost. Finally, we provide a detailed report of successful collaborative l
anguage model pretraining with nearly 50 participants.
****
Neural Ensemble Search for Uncertainty Estimation and Dataset Shift
Sheheryar Zaidi, Arber Zela, Thomas Elsken, Chris C Holmes, Frank Hutter, Yee Te
h
Ensembles of neural networks achieve superior performance compared to standalone
 networks in terms of accuracy, uncertainty calibration and robustness to datase
t shift. Deep ensembles, a state-of-the-art method for uncertainty estimation, o
nly ensemble random initializations of a fixed architecture. Instead, we propose
 two methods for automatically constructing ensembles with varying architectures
, which implicitly trade-off individual architectures’ strengths against the ens
emble’s diversity and exploit architectural variation as a source of diversity.
On a variety of classification tasks and modern architecture search spaces, we s
how that the resulting ensembles outperform deep ensembles not only in terms of
accuracy but also uncertainty calibration and robustness to dataset shift. Our f
urther analysis and ablation studies provide evidence of higher ensemble diversi
ty due to architectural variation, resulting in ensembles that can outperform de

ep ensembles, even when having weaker average base learners. To foster reproduci bility, our code is available: https://github.com/automl/nes
****
Finding Bipartite Components in Hypergraphs
Peter Macgregor, He Sun
Hypergraphs are important objects to model ternary or higher-order relations of objects, and have a number of applications in analysing many complex datasets oc curring in practice. In this work we study a new heat diffusion process in hype rgraphs, and employ this process to design a polynomial-time algorithm that appr oximately finds bipartite components in a hypergraph. We theoretically prove th e performance of our proposed algorithm, and compare it against the previous sta te-of-the-art through extensive experimental analysis on both synthetic and real -world datasets. We find that our new algorithm consistently and significantly o utperforms the previous state-of-the-art across a wide range of hypergraphs.
****
Hit and Lead Discovery with Explorative RL and Fragment-based Molecule Generatio n
Soojung Yang, Doyeong Hwang, Seul Lee, Seongok Ryu, Sung Ju Hwang
Recently, utilizing reinforcement learning (RL) to generate molecules with desir ed properties has been highlighted as a promising strategy for drug design. Mole cular docking program -- a physical simulation that estimates protein-small mole cule binding affinity -- can be an ideal reward scoring function for RL, as it i s a straightforward proxy of the therapeutic potential. Still, two imminent chal lenges exist for this task. First, the models often fail to generate chemically realistic and pharmacochemically acceptable molecules. Second, the docking score optimization is a difficult exploration problem that involves many local optima and less smooth surface with respect to molecular structure. To tackle these ch allenges, we propose a novel RL framework that generates pharmacochemically acce ptable molecules with large docking scores. Our method -- Fragment-based generat ive RL with Explorative Experience replay for Drug design (FREED) -- constrains the generated molecules to a realistic and qualified chemical space and effectiv ely explores the space to find drugs by coupling our fragment-based generation m ethod and a novel error-prioritized experience replay (PER). We also show that o ur model performs well on both de novo and scaffold-based schemes. Our model pro duces molecules of higher quality compared to existing methods while achieving s tate-of-the-art performance on two of three targets in terms of the docking scor es of the generated molecules. We further show with ablation studies that our me thod, predictive error-PER (FREED(PE)), significantly improves the model perform ance.
****
Proxy Convexity: A Unified Framework for the Analysis of Neural Networks Trained by Gradient Descent
Spencer Frei, Quanquan Gu
Although the optimization objectives for learning neural networks are highly non -convex, gradient-based methods have been wildly successful at learning neural n etworks in practice. This juxtaposition has led to a number of recent studies on provable guarantees for neural networks trained by gradient descent. Unfortunat ely, the techniques in these works are often highly specific to the particular s etup in each problem, making it difficult to generalize across different setting s. To address this drawback in the literature, we propose a unified non-convex o ptimization framework for the analysis of neural network training. We introduce the notions of proxy convexity and proxy Polyak-Lojasiewicz (PL) inequalities, w hich are satisfied if the original objective function induces a proxy objective function that is implicitly minimized when using gradient methods. We show that stochastic gradient descent (SGD) on objectives satisfying proxy convexity or th e proxy PL inequality leads to efficient guarantees for proxy objective function s. We further show that many existing guarantees for neural networks trained by gradient descent can be unified through proxy convexity and proxy PL inequalitie s.
****

Covariance-Aware Private Mean Estimation Without Private Covariance Estimation
Gavin Brown, Marco Gaboardi, Adam Smith, Jonathan Ullman, Lydia Zakynthinou
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Label consistency in overfitted generalized k-means
Linfan Zhang, 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.
****
Open-set Label Noise Can Improve Robustness Against Inherent Label Noise
Hongxin Wei, Lue Tao, RENCHUNZI XIE, Bo An
Learning with noisy labels is a practically challenging problem in weakly supervised learning. In the existing literature, open-set noises are always considered to be poisonous for generalization, similar to closed-set noises. In this paper, we empirically show that open-set noisy labels can be non-toxic and even benefit the robustness against inherent noisy labels. Inspired by the observations, we propose a simple yet effective regularization by introducing Open-set samples with Dynamic Noisy Labels (ODNL) into training. With ODNL, the extra capacity of the neural network can be largely consumed in a way that does not interfere with learning patterns from clean data. Through the lens of SGD noise, we show that the noises induced by our method are random-direction, conflict-free and biased, which may help the model converge to a flat minimum with superior stability and enforce the model to produce conservative predictions on Out-of-Distribution instances. Extensive experimental results on benchmark datasets with various types of noisy labels demonstrate that the proposed method not only enhances the performance of many existing robust algorithms but also achieves significant improvement on Out-of-Distribution detection tasks even in the label noise setting.
****
The Complexity of Sparse Tensor PCA
Davin Choo, Tommaso d'Orsi
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 Elect
Cem Anil, Xuchan Bao
Voting systems have a wide range of applications including recommender systems, web search, product design and elections. Limited by the lack of general-purpose analytical tools, it is difficult to hand-engineer desirable voting rules for each use case. For this reason, it is appealing to automatically discover voting rules geared towards each scenario. In this paper, we show that set-input neural network architectures such as Set Transformers, fully-connected graph networks and DeepSets are both theoretically and empirically well-suited for learning voting rules. In particular, we show that these network models can not only mimic a number of existing voting rules to compelling accuracy --- both position-based (such as Plurality and Borda) and comparison-based (such as Kemeny, Copeland and Maximin) --- but also discover near-optimal voting rules that maximize different social welfare functions. Furthermore, the learned voting rules generalize well to different voter utility distributions and election sizes unseen during training.
****
KALE Flow: A Relaxed KL Gradient Flow for Probabilities with Disjoint Support
Pierre Glaser, Michael Arbel, Arthur Gretton
We study the gradient flow for a relaxed approximation to the Kullback-Leibler (

KL) divergencebetween a moving source and a fixed target distribution.This appro
ximation, termed theKALE (KL approximate lower-bound estimator), solves a regula
rized version ofthe Fenchel dual problem defining the KL over a restricted class
 of functions.When using a Reproducing Kernel Hilbert Space (RKHS) to define the
 functionclass, we show that the KALE continuously interpolates between the KL a
nd theMaximum Mean Discrepancy (MMD). Like the MMD and other Integral Probabilit
yMetrics, the KALE remains well defined for mutually singulardistributions. None
theless, the KALE inherits from the limiting KL a greater sensitivity to mismatc
h in the support of the distributions, compared with the MMD. These two properti
es make theKALE gradient flow particularly well suited when the target distribut
ion is supported on a low-dimensional manifold. Under an assumption of sufficien
t smoothness of the trajectories, we show the global convergence of the KALE flo
w. We propose a particle implementation of the flow given initial samples from t
he source and the target distribution, which we use to empirically confirm the K
ALE's properties.
****
When Is Generalizable Reinforcement Learning Tractable?
Dhruv Malik, Yuanzhi Li, Pradeep Ravikumar
Agents trained by reinforcement learning (RL) often fail to generalize beyond th
e environment they were trained in, even when presented with new scenarios that
seem similar to the training environment. We study the query complexity required
 to train RL agents that generalize to multiple environments. Intuitively, tract
able generalization is only possible when the environments are similar or close
in some sense. To capture this, we introduce Weak Proximity, a natural structura
l condition that requires the environments to have highly similar transition and
 reward functions and share a policy providing optimal value. Despite such share
d structure, we prove that tractable generalization is impossible in the worst c
ase. This holds even when each individual environment can be efficiently solved
to obtain an optimal linear policy, and when the agent possesses a generative mo
del. Our lower bound applies to the more complex task of representation learning
 for efficient generalization to multiple environments. On the positive side, we
 introduce Strong Proximity, a strengthened condition which we prove is sufficie
nt for efficient generalization.
****
Relational Self-Attention: What's Missing in Attention for Video Understanding
Manjin Kim, Heeseung Kwon, CHUNYU WANG, Suha Kwak, Minsu Cho
Convolution has been arguably the most important feature transform for modern ne
ural networks, leading to the advance of deep learning. Recent emergence of Tra
nsformer networks, which replace convolution layers with self-attention blocks,
 has revealed the limitation of stationary convolution kernels and opened the do
or to the era of dynamic feature transforms. The existing dynamic transforms, in
cluding self-attention, however, are all limited for video understanding where c
orrespondence relations in space and time, i.e., motion information, are crucial
 for effective representation. In this work, we introduce a relational feature t
ransform, dubbed the relational self-attention (RSA), that leverages rich struct
ures of spatio-temporal relations in videos by dynamically generating relational
 kernels and aggregating relational contexts. Our experiments and ablation studi
es show that the RSA network substantially outperforms convolution and self-atte
ntion counterparts, achieving the state of the art on the standard motion-centri
c benchmarks for video action recognition, such as Something-Something-V1&V2, Di
ving48, and FineGym.
****
Towards Enabling Meta-Learning from Target Models
Su Lu, Han-Jia Ye, Le Gan, De-Chuan Zhan
Requests for name changes in the electronic proceedings will be accepted with 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 Near-Optimal Algorithm for Debiasing Trained Machine Learning Models

Ibrahim M. Alabdulmohsin, Mario Lucic
We present a scalable post-processing algorithm for debiasing trained models, in
cluding deep neural networks (DNNs), which we prove to be near-optimal by boundi
ng its excess Bayes risk. We empirically validate its advantages on standard be
nchmark datasets across both classical algorithms as well as modern DNN architec
tures and demonstrate that it outperforms previous post-processing methods while
 performing on par with in-processing. In addition, we show that the proposed al
gorithm is particularly effective for models trained at scale where post-process
ing is a natural and practical choice.
****
GENESIS-V2: Inferring Unordered Object Representations without Iterative Refinem
ent
Martin Engelcke, Oiwi Parker Jones, Ingmar Posner
Advances in unsupervised learning of object-representations have culminated in t
he development of a broad range of methods for unsupervised object segmentation
and interpretable object-centric scene generation. These methods, however, are l
imited to simulated and real-world datasets with limited visual complexity. More
over, object representations are often inferred using RNNs which do not scale we
ll to large images or iterative refinement which avoids imposing an unnatural or
dering on objects in an image but requires the a priori initialisation of a fixe
d number of object representations. In contrast to established paradigms, this w
ork proposes an embedding-based approach in which embeddings of pixels are clust
ered in a differentiable fashion using a stochastic stick-breaking process. Simi
lar to iterative refinement, this clustering procedure also leads to randomly or
dered object representations, but without the need of initialising a fixed numbe
r of clusters a priori. This is used to develop a new model, GENESIS-v2, which c
an infer a variable number of object representations without using RNNs or itera
tive refinement. We show that GENESIS-v2 performs strongly in comparison to rece
nt baselines in terms of unsupervised image segmentation and object-centric scen
e generation on established synthetic datasets as well as more complex real-worl
d datasets.
****
How Data Augmentation affects Optimization for Linear Regression
Boris Hanin, Yi Sun
Though data augmentation has rapidly emerged as a key tool for optimization in m
odern machine learning, a clear picture of how augmentation schedules affect opt
imization and interact with optimization hyperparameters such as learning rate i
s nascent. In the spirit of classical convex optimization and recent work on imp
licit bias, the present work analyzes the effect of augmentation on optimization
 in the simple convex setting of linear regression with MSE loss.We find joint s
chedules for learning rate and data augmentation scheme under which augmented gr
adient descent provably converges and characterize the resulting minimum. Our re
sults apply to arbitrary augmentation schemes, revealing complex interactions be
tween learning rates and augmentations even in the convex setting. Our approach
interprets augmented (S)GD as a stochastic optimization method for a time-varyin
g sequence of proxy losses. This gives a unified way to analyze learning rate, b
atch size, and augmentations ranging from additive noise to random projections.
From this perspective, our results, which also give rates of convergence, can be
 viewed as Monro-Robbins type conditions for augmented (S)GD.
****
An Exact Characterization of the Generalization Error for the Gibbs Algorithm
Gholamali Aminian, Yuheng Bu, Laura Toni, Miguel Rodrigues, Gregory Wornell
Various approaches have been developed to upper bound the generalization error o
f a supervised learning algorithm. However, existing bounds are often loose and
lack of guarantees. As a result, they may fail to characterize the exact general
ization ability of a learning algorithm.Our main contribution is an exact charac
terization of the expected generalization error of the well-known Gibbs algorith
m (a.k.a. Gibbs posterior) using symmetrized KL information between the input tr
aining samples and the output hypothesis. Our result can be applied to tighten e
xisting expected generalization error and PAC-Bayesian bounds. Our approach is v

ersatile, as it also characterizes the generalization error of the Gibbs algorithm with data-dependent regularizer and that of the Gibbs algorithm in the asymptotic regime, where it converges to the empirical risk minimization algorithm. Of particular relevance, our results highlight the role the symmetrized KL information plays in controlling the generalization error of the Gibbs algorithm.
****
Subgaussian and Differentiable Importance Sampling for Off-Policy Evaluation and Learning
Alberto Maria Metelli, Alessio Russo, Marcello Restelli
Importance Sampling (IS) is a widely used building block for a large variety of off-policy estimation and learning algorithms. However, empirical and theoretical studies have progressively shown that vanilla IS leads to poor estimations whenever the behavioral and target policies are too dissimilar. In this paper, we analyze the theoretical properties of the IS estimator by deriving a novel anticoncentration bound that formalizes the intuition behind its undesired behavior. Then, we propose a new class of IS transformations, based on the notion of power mean. To the best of our knowledge, the resulting estimator is the first to achieve, under certain conditions, two key properties: (i) it displays a subgaussian concentration rate; (ii) it preserves the differentiability in the target distribution. Finally, we provide numerical simulations on both synthetic examples and contextual bandits, in comparison with off-policy evaluation and learning baselines.
****
Rethinking gradient sparsification as total error minimization
Atal Sahu, Aritra Dutta, Ahmed M. Abdelmoniem, Trambak Banerjee, Marco Canini, Panos Kalnis
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Approximate optimization of convex functions with outlier noise
Anindya De, Sanjeev Khanna, Huan Li, MohammadHesam NikpeySalekde
We study the problem of minimizing a convex function given by a zeroth order oracle that is possibly corrupted by {\em outlier noise}. Specifically, we assume the function values at some points of the domain are corrupted arbitrarily by an adversary, with the only restriction being that the total volume of corrupted points is bounded. The goal then is to find a point close to the function's minimizer using access to the corrupted oracle.We first prove a lower bound result showing that, somewhat surprisingly, one cannot hope to approximate the minimizer {\em nearly as well} as one might expect, even if one is allowed {\em an unbounded number} of queries to the oracle. Complementing this negative result, we then develop an efficient algorithm that outputs a point close to the minimizer of the convex function, where the specific distance matches {\em exactly}, up to constant factors, the distance bound shown in our lower bound result.
****
Fair Classification with Adversarial Perturbations
L. Elisa Celis, Anay Mehrotra, Nisheeth Vishnoi
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Distributed Saddle-Point Problems Under Data Similarity
Aleksandr Beznosikov, Gesualdo Scutari, Alexander Rogozin, Alexander Gasnikov
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

Combining Latent Space and Structured Kernels for Bayesian Optimization over Com binatorial Spaces
Aryan Deshwal, Jana Doppa
We consider the problem of optimizing combinatorial spaces (e.g., sequences, tre es, and graphs) using expensive black-box function evaluations. For example, opt imizing molecules for drug design using physical lab experiments. Bayesian optim ization (BO) is an efficient framework for solving such problems by intelligentl y selecting the inputs with high utility guided by a learned surrogate model. A recent BO approach for combinatorial spaces is through a reduction to BO over co ntinuous spaces by learning a latent representation of structures using deep gen erative models (DGMs). The selected input from the continuous space is decoded i nto a discrete structure for performing function evaluation. However, the surrog ate model over the latent space only uses the information learned by the DGM, wh ich may not have the desired inductive bias to approximate the target black-box function. To overcome this drawback, this paper proposes a principled approach r eferred as LADDER. The key idea is to define a novel structure-coupled kernel th at explicitly integrates the structural information from decoded structures with the learned latent space representation for better surrogate modeling. Our expe riments on real-world benchmarks show that LADDER significantly improves over th e BO over latent space method, and performs better or similar to state-of-the-ar t methods.
****
Gradual Domain Adaptation without Indexed Intermediate Domains
Hong-You Chen, Wei-Lun Chao
The effectiveness of unsupervised domain adaptation degrades when there is a lar ge discrepancy between the source and target domains. Gradual domain adaption (G DA) is one promising way to mitigate such an issue, by leveraging additional unl abeled data that gradually shift from the source to the target. Through sequenti ally adapting the model along the "indexed" intermediate domains, GDA substantia lly improves the overall adaptation performance. In practice, however, the extra unlabeled data may not be separated into intermediate domains and indexed prope rly, limiting the applicability of GDA. In this paper, we investigate how to dis cover the sequence of intermediate domains when it is not already available. Con cretely, we propose a coarse-to-fine framework, which starts with a coarse domai n discovery step via progressive domain discriminator training. This coarse doma in sequence then undergoes a fine indexing step via a novel cycle-consistency lo ss, which encourages the next intermediate domain to preserve sufficient discrim inative knowledge of the current intermediate domain. The resulting domain seque nce can then be used by a GDA algorithm. On benchmark data sets of GDA, we show that our approach, which we name Intermediate DOmain Labeler (IDOL), can lead to comparable or even better adaptation performance compared to the pre-defined do main sequence, making GDA more applicable and robust to the quality of domain se quences. Codes are available at https://github.com/hongyouc/IDOL.
****
K-level Reasoning for Zero-Shot Coordination in Hanabi
Brandon Cui, Hengyuan Hu, Luis Pineda, Jakob Foerster
Requests for name changes in the electronic proceedings will be accepted with 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 Markov State Abstractions for Deep Reinforcement Learning
Cameron Allen, Neev Parikh, Omer Gottesman, George Konidaris
A fundamental assumption of reinforcement learning in Markov decision processes (MDPs) is that the relevant decision process is, in fact, Markov. However, when MDPs have rich observations, agents typically learn by way of an abstract state representation, and such representations are not guaranteed to preserve the Mark ov property. We introduce a novel set of conditions and prove that they are suff icient for learning a Markov abstract state representation. We then describe a p ractical training procedure that combines inverse model estimation and temporal

contrastive learning to learn an abstraction that approximately satisfies these conditions. Our novel training objective is compatible with both online and offline training: it does not require a reward signal, but agents can capitalize on reward information when available. We empirically evaluate our approach on a visual gridworld domain and a set of continuous control benchmarks. Our approach learns representations that capture the underlying structure of the domain and lead to improved sample efficiency over state-of-the-art deep reinforcement learning with visual features---often matching or exceeding the performance achieved with hand-designed compact state information.
****
Towards Deeper Deep Reinforcement Learning with Spectral Normalization
Nils Bjorck, Carla P. Gomes, Kilian Q. Weinberger
In computer vision and natural language processing, innovations in model architecture that increase model capacity have reliably translated into gains in performance. In stark contrast with this trend, state-of-the-art reinforcement learning (RL) algorithms often use small MLPs, and gains in performance typically originate from algorithmic innovations. It is natural to hypothesize that small datasets in RL necessitate simple models to avoid overfitting; however, this hypothesis is untested. In this paper we investigate how RL agents are affected by exchanging the small MLPs with larger modern networks with skip connections and normalization, focusing specifically on actor-critic algorithms. We empirically verify that naively adopting such architectures leads to instabilities and poor performance, likely contributing to the popularity of simple models in practice. However, we show that dataset size is not the limiting factor, and instead argue that instability from taking gradients through the critic is the culprit. We demonstrate that spectral normalization (SN) can mitigate this issue and enable stable training with large modern architectures. After smoothing with SN, larger models yield significant performance improvements --- suggesting that more ``easy'' gains may be had by focusing on model architectures in addition to algorithmic innovations.
****
Functionally Regionalized Knowledge Transfer for Low-resource Drug Discovery
Huaxiu Yao, Ying Wei, Long-Kai Huang, Ding Xue, Junzhou Huang, Zhenhui (Jessie) Li
More recently, there has been a surge of interest in employing machine learning approaches to expedite the drug discovery process where virtual screening for hit discovery and ADMET prediction for lead optimization play essential roles. One of the main obstacles to the wide success of machine learning approaches in these two tasks is that the number of compounds labeled with activities or ADMET properties is too small to build an effective predictive model. This paper seeks to remedy the problem by transferring the knowledge from previous assays, namely in-vivo experiments, by different laboratories and against various target proteins. To accommodate these wildly different assays and capture the similarity between assays, we propose a functional rationalized meta-learning algorithm FRML for such knowledge transfer. FRML constructs the predictive model with layers of neural sub-networks or so-called functional regions. Building on this, FRML shares an initialization for the weights of the predictive model across all assays, while customizes it to each assay with a region localization network choosing the pertinent regions. The compositionality of the model improves the capacity of generalization to various and even out-of-distribution tasks. Empirical results on both virtual screening and ADMET prediction validate the superiority of FRML over state-of-the-art baselines powered with interpretability in assay relationship.
****
Memory-Efficient Approximation Algorithms for Max-k-Cut and Correlation Clustering
Nimita Shinde, Vishnu Narayanan, James Saunderson
Requests for name changes in the electronic proceedings will be accepted with 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.
****
Panoptic 3D Scene Reconstruction From a Single RGB Image
Manuel Dahnert, Ji Hou, Matthias Niessner, Angela Dai
Richly segmented 3D scene reconstructions are an integral basis for many high-level scene understanding tasks, such as for robotics, motion planning, or augmented reality. Existing works in 3D perception from a single RGB image tend to focus on geometric reconstruction only, or geometric reconstruction with semantic segmentation or instance segmentation.Inspired by 2D panoptic segmentation, we propose to unify the tasks of geometric reconstruction, 3D semantic segmentation, and 3D instance segmentation into the task of panoptic 3D scene reconstruction -- from a single RGB image, predicting the complete geometric reconstruction of the scene in the camera frustum of the image, along with semantic and instance segmentations.We propose a new approach for holistic 3D scene understanding from a single RGB image which learns to lift and propagate 2D features from an input image to a 3D volumetric scene representation.Our panoptic 3D reconstruction metric evaluates both geometric reconstruction quality as well as panoptic segmentation.Our experiments demonstrate that our approach for panoptic 3D scene reconstruction outperforms alternative approaches for this task.
****
Measuring Generalization with Optimal Transport
Ching-Yao Chuang, Youssef Mroueh, Kristjan Greenewald, Antonio Torralba, Stefanie Jegelka
Understanding the generalization of deep neural networks is one of the most important tasks in deep learning. Although much progress has been made, theoretical error bounds still often behave disparately from empirical observations. In this work, we develop margin-based generalization bounds, where the margins are normalized with optimal transport costs between independent random subsets sampled from the training distribution. In particular, the optimal transport cost can be interpreted as a generalization of variance which captures the structural properties of the learned feature space. Our bounds robustly predict the generalization error, given training data and network parameters, on large scale datasets. Theoretically, we demonstrate that the concentration and separation of features play crucial roles in generalization, supporting empirical results in the literature.
****
Uniform Concentration Bounds toward a Unified Framework for Robust Clustering
Debolina Paul, Saptarshi Chakraborty, Swagatam Das, Jason Xu
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Learning Signal-Agnostic Manifolds of Neural Fields
Yilun Du, Katie Collins, Josh Tenenbaum, Vincent Sitzmann
Deep neural networks have been used widely to learn the latent structure of datasets, across modalities such as images, shapes, and audio signals. However, existing models are generally modality-dependent, requiring custom architectures and objectives to process different classes of signals. We leverage neural fields to capture the underlying structure in image, shape, audio and cross-modal audiovisual domains in a modality-independent manner. We cast our task as one of learning a manifold, where we aim to infer a low-dimensional, locally linear subspace in which our data resides. By enforcing coverage of the manifold, local linearity, and local isometry, our model -- dubbed GEM -- learns to capture the underlying structure of datasets across modalities. We can then travel along linear regions of our manifold to obtain perceptually consistent interpolations between samples, and can further use GEM to recover points on our manifold and glean not only diverse completions of input images, but cross-modal hallucinations of audio or image signals. Finally, we show that by walking across the underlying manifold of GEM, we may generate new samples in our signal domains.

****
Low-dimensional Structure in the Space of Language Representations is Reflected in Brain Responses
Richard Antonello, Javier S. Turek, Vy Vo, Alexander Huth
How related are the representations learned by neural language models, translation models, and language tagging tasks? We answer this question by adapting an encoder-decoder transfer learning method from computer vision to investigate the structure among 100 different feature spaces extracted from hidden representations of various networks trained on language tasks.This method reveals a low-dimensional structure where language models and translation models smoothly interpolate between word embeddings, syntactic and semantic tasks, and future word embeddings. We call this low-dimensional structure a language representation embedding because it encodes the relationships between representations needed to process language for a variety of NLP tasks. We find that this representation embedding can predict how well each individual feature space maps to human brain responses to natural language stimuli recorded using fMRI. Additionally, we find that the principal dimension of this structure can be used to create a metric which highlights the brain's natural language processing hierarchy. This suggests that the embedding captures some part of the brain's natural language representation structure.
****
On the Suboptimality of Thompson Sampling in High Dimensions
Raymond Zhang, Richard Combes
In this paper we consider Thompson Sampling for combinatorial semi-bandits. We demonstrate that, perhaps surprisingly, Thompson Sampling is sub-optimal for this problem in the sense that its regret scales exponentially in the ambient dimension, and its minimax regret scales almost linearly. This phenomenon occurs under a wide variety of assumptions including both non-linear and linear reward functions in the Bernoulli distribution setting. We also show that including a fixed amount of forced exploration to Thompson Sampling does not alleviate the problem. We complement our theoretical results with numerical results and show that in practice Thompson Sampling indeed can perform very poorly in some high dimension situations.
****
Learning Debiased and Disentangled Representations for Semantic Segmentation
Sanghyeok Chu, Dongwan Kim, Bohyung Han
Deep neural networks are susceptible to learn biased models with entangled feature representations, which may lead to subpar performances on various downstream tasks. This is particularly true for under-represented classes, where a lack of diversity in the data exacerbates the tendency. This limitation has been addressed mostly in classification tasks, but there is little study on additional challenges that may appear in more complex dense prediction problems including semantic segmentation. To this end, we propose a model-agnostic and stochastic training scheme for semantic segmentation, which facilitates the learning of debiased and disentangled representations. For each class, we first extract class-specific information from the highly entangled feature map. Then, information related to a randomly sampled class is suppressed by a feature selection process in the feature space. By randomly eliminating certain class information in each training iteration, we effectively reduce feature dependencies among classes, and the model is able to learn more debiased and disentangled feature representations. Models trained with our approach demonstrate strong results on multiple semantic segmentation benchmarks, with especially notable performance gains on under-represented classes.
****
Diversity Matters When Learning From Ensembles
Giung Nam, Jongmin Yoon, Yoonho Lee, Juho Lee
Deep ensembles excel in large-scale image classification tasks both in terms of prediction accuracy and calibration. Despite being simple to train, the computation and memory cost of deep ensembles limits their practicability. While some recent works propose to distill an ensemble model into a single model to reduce su

ch costs, there is still a performance gap between the ensemble and distilled mo
dels. We propose a simple approach for reducing this gap, i.e., making the disti
lled performance close to the full ensemble. Our key assumption is that a distil
led model should absorb as much function diversity inside the ensemble as possib
le. We first empirically show that the typical distillation procedure does not e
ffectively transfer such diversity, especially for complex models that achieve n
ear-zero training error. To fix this, we propose a perturbation strategy for dis
tillation that reveals diversity by seeking inputs for which ensemble member out
puts disagree. We empirically show that a model distilled with such perturbed sa
mples indeed exhibits enhanced diversity, leading to improved performance.
****
Locally Valid and Discriminative Prediction Intervals for Deep Learning Models
Zhen Lin, Shubhendu Trivedi, Jimeng Sun
Crucial for building trust in deep learning models for critical real-world appli
cations is efficient and theoretically sound uncertainty quantification, a task
that continues to be challenging. Useful uncertainty information is expected to
have two key properties: It should be valid (guaranteeing coverage) and discrimi
native (more uncertain when the expected risk is high). Moreover, when combined
with deep learning (DL) methods, it should be scalable and affect the DL model p
erformance minimally. Most existing Bayesian methods lack frequentist coverage g
uarantees and usually affect model performance. The few available frequentist me
thods are rarely discriminative and/or violate coverage guarantees due to unreal
istic assumptions. Moreover, many methods are expensive or require substantial m
odifications to the base neural network. Building upon recent advances in confor
mal prediction [13, 33] and leveraging the classical idea of kernel regression,
we propose Locally Valid and Discriminative prediction intervals (LVD), a simple
, efficient, and lightweight method to construct discriminative prediction inter
vals (PIs) for almost any DL model. With no assumptions on the data distribution
, such PIs also offer finite-sample local coverage guarantees (contrasted to the
 simpler marginal coverage). We empirically verify, using diverse datasets, that
 besides being the only locally valid method for DL, LVD also exceeds or matches
 the performance (including coverage rate and prediction accuracy) of existing u
ncertainty quantification methods, while offering additional benefits in scalabi
lity and flexibility.
****
Personalized Federated Learning With Gaussian Processes
Idan Achituve, Aviv Shamsian, Aviv Navon, Gal Chechik, Ethan Fetaya
Federated learning aims to learn a global model that performs well on client dev
ices with limited cross-client communication. Personalized federated learning (P
FL) further extends this setup to handle data heterogeneity between clients by l
earning personalized models. A key challenge in this setting is to learn effecti
vely across clients even though each client has unique data that is often limite
d in size. Here we present pFedGP, a solution to PFL that is based on Gaussian p
rocesses (GPs) with deep kernel learning. GPs are highly expressive models that
work well in the low data regime due to their Bayesian nature.However, applying
GPs to PFL raises multiple challenges. Mainly, GPs performance depends heavily o
n access to a good kernel function, and learning a kernel requires a large train
ing set. Therefore, we propose learning a shared kernel function across all clie
nts, parameterized by a neural network, with a personal GP classifier for each c
lient. We further extend pFedGP to include inducing points using two novel metho
ds, the first helps to improve generalization in the low data regime and the sec
ond reduces the computational cost. We derive a PAC-Bayes generalization bound o
n novel clients and empirically show that it gives non-vacuous guarantees. Exten
sive experiments on standard PFL benchmarks with CIFAR-10, CIFAR-100, and CINIC-
10, and on a new setup of learning under input noise show that pFedGP achieves w
ell-calibrated predictions while significantly outperforming baseline methods, r
eaching up to 21% in accuracy gain.
****
Risk Bounds for Over-parameterized Maximum Margin Classification on Sub-Gaussian
 Mixtures

Yuan Cao, Quanquan Gu, Mikhail Belkin
Modern machine learning systems such as deep neural networks are often highly over-parameterized so that they can fit the noisy training data exactly, yet they can still achieve small test errors in practice. In this paper, we study this "benign overfitting" phenomenon of the maximum margin classifier for linear classification problems. Specifically, we consider data generated from sub-Gaussian mixtures, and provide a tight risk bound for the maximum margin linear classifier in the over-parameterized setting. Our results precisely characterize the condition under which benign overfitting can occur in linear classification problems, and improve on previous work. They also have direct implications for over-parameterized logistic regression.
****
Implicit SVD for Graph Representation Learning
Sami Abu-El-Haija, Hesham Mostafa, Marcel Nassar, Valentino Crespi, Greg Ver Steeg, Aram Galstyan
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Offline Model-based Adaptable Policy Learning
Xiong-Hui Chen, Yang Yu, Qingyang Li, Fan-Ming Luo, Zhiwei Qin, Wenjie Shang, Jieping Ye
In reinforcement learning, a promising direction to avoid online trial-and-error costs is learning from an offline dataset. Current offline reinforcement learning methods commonly learn in the policy space constrained to in-support regions by the offline dataset, in order to ensure the robustness of the outcome policies. Such constraints, however, also limit the potential of the outcome policies. In this paper, to release the potential of offline policy learning, we investigate the decision-making problems in out-of-support regions directly and propose offline Model-based Adaptable Policy LEarning (MAPLE). By this approach, instead of learning in in-support regions, we learn an adaptable policy that can adapt its behavior in out-of-support regions when deployed. We conduct experiments on MuJoCo controlling tasks with offline datasets. The results show that the proposed method can make robust decisions in out-of-support regions and achieve better performance than SOTA algorithms.
****
Multilingual Pre-training with Universal Dependency Learning
Kailai Sun, Zuchao Li, Hai Zhao
The pre-trained language model (PrLM) demonstrates domination in downstream natural language processing tasks, in which multilingual PrLM takes advantage of language universality to alleviate the issue of limited resources for low-resource languages. Despite its successes, the performance of multilingual PrLM is still unsatisfactory, when multilingual PrLMs only focus on plain text and ignore obvious universal linguistic structure clues. Existing PrLMs have shown that monolingual linguistic structure knowledge may bring about better performance. Thus we propose a novel multilingual PrLM that supports both explicit universal dependency parsing and implicit language modeling. Syntax in terms of universal dependency parse serves as not only pre-training objective but also learned representation in our model, which brings unprecedented PrLM interpretability and convenience in downstream task use. Our model outperforms two popular multilingual PrLM, multilingual-BERT and XLM-R, on cross-lingual natural language understanding (NLU) benchmarks and linguistic structure parsing datasets, demonstrating the effectiveness and stronger cross-lingual modeling capabilities of our approach.
****
Parameter-free HE-friendly Logistic Regression
Junyoung Byun, Woojin Lee, Jaewook Lee
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-auth

ors prior to requesting a name change in the electronic proceedings.
****
Active clustering for labeling training data
Quentin Lutz, Elie de Panafieu, Maya Stein, Alex Scott
Gathering training data is a key step of any supervised learning task, and it is
 both critical and expensive. Critical, because the quantity and quality of the
training data has a high impact on the performance of the learned function. Expe
nsive, because most practical cases rely on humans-in-the-loop to label the data
. The process of determining the correct labels is much more expensive than comp
aring two items to see whether they belong to the same class. Thus motivated, we
 propose a setting for training data gathering where the human experts perform t
he comparatively cheap task of answering pairwise queries, and the computer grou
ps the items into classes (which can be labeled cheaply at the very end of the p
rocess). Given the items, we consider two random models for the classes: one whe
re the set partition they form is drawn uniformly, the other one where each item
 chooses its class independently following a fixed distribution. In the first mo
del, we characterize the algorithms that minimize the average number of queries
required to cluster the items and analyze their complexity. In the second model,
 we analyze a specific algorithm family, propose as a conjecture that they reach
 the minimum average number of queries and compare their performance to a random
 approach. We also propose solutions to handle errors or inconsistencies in the
experts' answers.
****
Exploring Social Posterior Collapse in Variational Autoencoder for Interaction M
odeling
Chen Tang, Wei Zhan, Masayoshi Tomizuka
Multi-agent behavior modeling and trajectory forecasting are crucial for the saf
e navigation of autonomous agents in interactive scenarios. Variational Autoenco
der (VAE) has been widely applied in multi-agent interaction modeling to generat
e diverse behavior and learn a low-dimensional representation for interacting sy
stems. However, existing literature did not formally discuss if a VAE-based mode
l can properly encode interaction into its latent space. In this work, we argue
that one of the typical formulations of VAEs in multi-agent modeling suffers fro
m an issue we refer to as social posterior collapse, i.e., the model is prone to
 ignoring historical social context when predicting the future trajectory of an
agent. It could cause significant prediction errors and poor generalization perf
ormance. We analyze the reason behind this under-explored phenomenon and propose
 several measures to tackle it. Afterward, we implement the proposed framework a
nd experiment on real-world datasets for multi-agent trajectory prediction. In p
articular, we propose a novel sparse graph attention message-passing (sparse-GAM
P) layer, which helps us detect social posterior collapse in our experiments. In
 the experiments, we verify that social posterior collapse indeed occurs. Also,
the proposed measures are effective in alleviating the issue. As a result, the m
odel attains better generalization performance when historical social context is
 informative for prediction.
****
Ensembling Graph Predictions for AMR Parsing
Thanh Lam Hoang, Gabriele Picco, Yufang Hou, Young-Suk Lee, Lam Nguyen, Dzung Ph
an, Vanessa Lopez, Ramon Fernandez Astudillo
In many machine learning tasks, models are trained to predict structure data suc
h as graphs. For example, in natural language processing, it is very common to p
arse texts into dependency trees or abstract meaning representation (AMR) graphs
. On the other hand, ensemble methods combine predictions from multiple models t
o create a new one that is more robust and accurate than individual predictions.
 In the literature, there are many ensembling techniques proposed for classifica
tion or regression problems, however, ensemble graph prediction has not been stu
died thoroughly. In this work, we formalize this problem as mining the largest g
raph that is the most supported by a collection of graph predictions. As the pro
blem is NP-Hard, we propose an efficient heuristic algorithm to approximate the
optimal solution. To validate our approach, we carried out experiments in AMR pa

rsing problems. The experimental results demonstrate that the proposed approach can combine the strength of state-of-the-art AMR parsers to create new predictions that are more accurate than any individual models in five standard benchmark datasets.
****
On the interplay between data structure and loss function in classification problems
Stéphane d'Ascoli, Marylou Gabrié, Levent Sagun, Giulio Biroli
One of the central features of modern machine learning models, including deep neural networks, is their generalization ability on structured data in the over-parametrized regime. In this work, we consider an analytically solvable setup to investigate how properties of data impact learning in classification problems, and compare the results obtained for quadratic loss and logistic loss. Using methods from statistical physics, we obtain a precise asymptotic expression for the train and test errors of random feature models trained on a simple model of structured data. The input covariance is built from independent blocks allowing us to tune the saliency of low-dimensional structures and their alignment with respect to the target function.Our results show in particular that in the over-parametrized regime, the impact of data structure on both train and test error curves is greater for logistic loss than for mean-squared loss: the easier the task, the wider the gap in performance between the two losses at the advantage of the logistic. Numerical experiments on MNIST and CIFAR10 confirm our insights.
****
Near-optimal Offline and Streaming Algorithms for Learning Non-Linear Dynamical Systems
Suhas Kowshik, Dheeraj Nagaraj, 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.
****
Mixture Proportion Estimation and PU Learning:A Modern Approach
Saurabh Garg, Yifan Wu, Alexander J. Smola, 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-authors prior to requesting a name change in the electronic proceedings.
****
Escape saddle points by a simple gradient-descent based algorithm
Chenyi Zhang, Tongyang 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.
****
AC/DC: Alternating Compressed/DeCompressed Training of Deep Neural Networks
Alexandra Peste, Eugenia Iofinova, Adrian Vladu, Dan Alistarh
The increasing computational requirements of deep neural networks (DNNs) have led to significant interest in obtaining DNN models that are sparse, yet accurate. Recent work has investigated the even harder case of sparse training, where the DNN weights are, for as much as possible, already sparse to reduce computational costs during training. Existing sparse training methods are often empirical and can have lower accuracy relative to the dense baseline. In this paper, we present a general approach called Alternating Compressed/DeCompressed (AC/DC) training of DNNs, demonstrate convergence for a variant of the algorithm, and show that AC/DC outperforms existing sparse training methods in accuracy at similar computational budgets; at high sparsity levels, AC/DC even outperforms existing methods that rely on accurate pre-trained dense models. An important property of AC/DC is that it allows co-training of dense and sparse models, yielding accurate

sparse-dense model pairs at the end of the training process. This is useful in p
ractice, where compressed variants may be desirable for deployment in resource-c
onstrained settings without re-doing the entire training flow, and also provides
 us with insights into the accuracy gap between dense and compressed models.
****
HyperSPNs: Compact and Expressive Probabilistic Circuits
Andy Shih, Dorsa Sadigh, Stefano Ermon
Probabilistic circuits (PCs) are a family of generative models which allows for
the computation of exact likelihoods and marginals of its probability distributi
ons. PCs are both expressive and tractable, and serve as popular choices for dis
crete density estimation tasks. However, large PCs are susceptible to overfittin
g, and only a few regularization strategies (e.g., dropout, weight-decay) have b
een explored. We propose HyperSPNs: a new paradigm of generating the mixture wei
ghts of large PCs using a small-scale neural network. Our framework can be viewe
d as a soft weight-sharing strategy, which combines the greater expressiveness o
f large models with the better generalization and memory-footprint properties of
 small models. We show the merits of our regularization strategy on two state-o
f-the-art PC families introduced in recent literature -- RAT-SPNs and EiNETs --
and demonstrate generalization improvements in both models on a suite of density
 estimation benchmarks in both discrete and continuous domains.
****
Scaling Vision with Sparse Mixture of Experts
Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatto
n, André Susano Pinto, Daniel Keysers, Neil Houlsby
Sparsely-gated Mixture of Experts networks (MoEs) have demonstrated excellent sc
alability in Natural Language Processing. In Computer Vision, however, almost al
l performant networks are "dense", that is, every input is processed by every pa
rameter. We present a Vision MoE (V-MoE), a sparse version of the Vision Transfo
rmer, that is scalable and competitive with the largest dense networks. When app
lied to image recognition, V-MoE matches the performance of state-of-the-art net
works, while requiring as little as half of the compute at inference time. Furth
er, we propose an extension to the routing algorithm that can prioritize subsets
 of each input across the entire batch, leading to adaptive per-image compute. T
his allows V-MoE to trade-off performance and compute smoothly at test-time. Fin
ally, we demonstrate the potential of V-MoE to scale vision models, and train a
15B parameter model that attains 90.35% on ImageNet.
****
Two-sided fairness in rankings via Lorenz dominance
Virginie Do, Sam Corbett-Davies, Jamal Atif, Nicolas Usunier
We consider the problem of generating rankings that are fair towards both users
and item producers in recommender systems. We address both usual recommendation
(e.g., of music or movies) and reciprocal recommendation (e.g., dating). Followi
ng concepts of distributive justice in welfare economics, our notion of fairness
 aims at increasing the utility of the worse-off individuals, which we formalize
 using the criterion of Lorenz efficiency. It guarantees that rankings are Paret
o efficient, and that they maximally redistribute utility from better-off to wor
se-off, at a given level of overall utility. We propose to generate rankings by
maximizing concave welfare functions, and develop an efficient inference procedu
re based on the Frank-Wolfe algorithm. We prove that unlike existing approaches
based on fairness constraints, our approach always produces fair rankings. Our e
xperiments also show that it increases the utility of the worse-off at lower cos
ts in terms of overall utility.
****
Stability & Generalisation of Gradient Descent for Shallow Neural Networks witho
ut the Neural Tangent Kernel
Dominic Richards, Ilja Kuzborskij
We revisit on-average algorithmic stability of Gradient Descent (GD) for trainin
g overparameterised shallow neural networks and prove new generalisation
and excess risk bounds without the Neural Tangent Kernel (NTK) or Pol
yak-■ojasiewicz (PL) assumptions. In particular, we show oracle type bounds whic

h reveal that the generalisation and excess risk of GD is controlled by an inter
polating network with the shortest GD path from initialisation (in a sense, an i
nterpolating network with the smallest relative norm). While this was known for
 kernelised interpolants, our proof applies directly to networks trained by GD w
ithout intermediate kernelisation. At the same time, by relaxing oracle inequali
ties developed here we recover existing NTK-based risk bounds in a straightforwa
rd way, which demonstrates that our analysis is tighter. Finally, unlike most of
 the NTK-based analyses we focus on regression with label noise and show that GD
 with early stopping is consistent
****
Adversarial Intrinsic Motivation for Reinforcement Learning
Ishan Durugkar, Mauricio Tec, Scott Niekum, Peter Stone
Learning with an objective to minimize the mismatch with a reference distributio
n has been shown to be useful for generative modeling and imitation learning. In
 this paper, we investigate whether one such objective, the Wasserstein-1 distan
ce between a policy's state visitation distribution and a target distribution, c
an be utilized effectively for reinforcement learning (RL) tasks. Specifically,
this paper focuses on goal-conditioned reinforcement learning where the idealize
d (unachievable) target distribution has full measure at the goal. This paper in
troduces a quasimetric specific to Markov Decision Processes (MDPs) and uses thi
s quasimetric to estimate the above Wasserstein-1 distance. It further shows tha
t the policy that minimizes this Wasserstein-1 distance is the policy that reach
es the goal in as few steps as possible. Our approach, termed Adversarial Intrin
sic Motivation (AIM), estimates this Wasserstein-1 distance through its dual obj
ective and uses it to compute a supplemental reward function. Our experiments sh
ow that this reward function changes smoothly with respect to transitions in the
 MDP and directs the agent's exploration to find the goal efficiently. Additiona
lly, we combine AIM with Hindsight Experience Replay (HER) and show that the res
ulting algorithm accelerates learning significantly on several simulated robotic
s tasks when compared to other rewards that encourage exploration or accelerate
learning.
****
Machine Learning for Variance Reduction in Online Experiments
Yongyi Guo, Dominic Coey, Mikael Konutgan, Wenting Li, Chris Schoener, Matt Gold
man
Requests for name changes in the electronic proceedings will be accepted with 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.
****
L2ight: Enabling On-Chip Learning for Optical Neural Networks via Efficient in-s
itu Subspace Optimization
Jiaqi Gu, Hanqing Zhu, Chenghao Feng, Zixuan Jiang, Ray Chen, David Pan
Silicon-photonics-based optical neural network (ONN) is a promising hardware pla
tform that could represent a paradigm shift in efficient AI with its CMOS-compat
ibility, flexibility, ultra-low execution latency, and high energy efficiency. I
n-situ training on the online programmable photonic chips is appealing but still
 encounters challenging issues in on-chip implementability, scalability, and eff
iciency. In this work, we propose a closed-loop ONN on-chip learning framework L
2ight to enable scalable ONN mapping and efficient in-situ learning. L2ight adop
ts a three-stage learning flow that first calibrates the complicated photonic ci
rcuit states under challenging physical constraints, then performs photonic core
 mapping via combined analytical solving and zeroth-order optimization. A subspa
ce learning procedure with multi-level sparsity is integrated into L2ight to ena
ble in-situ gradient evaluation and fast adaptation, unleashing the power of opt
ics for real on-chip intelligence. Extensive experiments demonstrate our propose
d L2ight outperforms prior ONN training protocols with 3-order-of-magnitude high
er scalability and over 30x better efficiency, when benchmarked on various model
s and learning tasks. This synergistic framework is the first scalable on-chip l
earning solution that pushes this emerging field from intractable to scalable an

d further to efficient for next-generation self-learnable photonic neural chips. From a co-design perspective, L2ight also provides essential insights for hardware-restricted unitary subspace optimization and efficient sparse training. We open-source our framework at the link.
****
Towards Gradient-based Bilevel Optimization with Non-convex Followers and Beyond
Risheng Liu, Yaohua Liu, Shangzhi Zeng, Jin Zhang
In recent years, Bi-Level Optimization (BLO) techniques have received extensive attentions from both learning and vision communities. A variety of BLO models in complex and practical tasks are of non-convex follower structure in nature (a.k.a., without Lower-Level Convexity, LLC for short). However, this challenging class of BLOs is lack of developments on both efficient solution strategies and solid theoretical guarantees. In this work, we propose a new algorithmic framework, named Initialization Auxiliary and Pessimistic Trajectory Truncated Gradient Method (IAPTT-GM), to partially address the above issues. In particular, by introducing an auxiliary as initialization to guide the optimization dynamics and designing a pessimistic trajectory truncation operation, we construct a reliable approximate version of the original BLO in the absence of LLC hypothesis. Our theoretical investigations establish the convergence of solutions returned by IAPTT-GM towards those of the original BLO without LLC. As an additional bonus, we also theoretically justify the quality of our IAPTT-GM embedded with Nesterov's accelerated dynamics under LLC. The experimental results confirm both the convergence of our algorithm without LLC, and the theoretical findings under LLC.
****
Multi-Facet Clustering Variational Autoencoders
Fabian Falck, Haoting Zhang, Matthew Willetts, George Nicholson, Christopher Yau, Chris C Holmes
Work in deep clustering focuses on finding a single partition of data. However, high-dimensional data, such as images, typically feature multiple interesting characteristics one could cluster over. For example, images of objects against a background could be clustered over the shape of the object and separately by the colour of the background. In this paper, we introduce Multi-Facet Clustering Variational Autoencoders (MFCVAE), a novel class of variational autoencoders with a hierarchy of latent variables, each with a Mixture-of-Gaussians prior, that learns multiple clusterings simultaneously, and is trained fully unsupervised and end-to-end. MFCVAE uses a progressively-trained ladder architecture which leads to highly stable performance. We provide novel theoretical results for optimising the ELBO analytically with respect to the categorical variational posterior distribution, correcting earlier influential theoretical work. On image benchmarks, we demonstrate that our approach separates out and clusters over different aspects of the data in a disentangled manner. We also show other advantages of our model: the compositionality of its latent space and that it provides controlled generation of samples.
****
Synthetic Design: An Optimization Approach to Experimental Design with Synthetic Controls
Nick Doudchenko, Khashayar Khosravi, Jean Pouget-Abadie, Sébastien Lahaie, Miles Lubin, Vahab Mirrokni, Jann Spiess, guido imbens
We investigate the optimal design of experimental studies that have pre-treatment outcome data available. The average treatment effect is estimated as the difference between the weighted average outcomes of the treated and control units. A number of commonly used approaches fit this formulation, including the difference-in-means estimator and a variety of synthetic-control techniques. We propose several methods for choosing the set of treated units in conjunction with the weights. Observing the NP-hardness of the problem, we introduce a mixed-integer programming formulation which selects both the treatment and control sets and unit weightings. We prove that these proposed approaches lead to qualitatively different experimental units being selected for treatment. We use simulations based on publicly available data from the US Bureau of Labor Statistics that show improvements in terms of mean squared error and statistical power when compared to si

mple and commonly used alternatives such as randomized trials.
****
Ranking Policy Decisions
Hadrien Pouget, Hana Chockler, Youcheng Sun, Daniel Kroening
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Searching the Search Space of Vision Transformer
Minghao Chen, Kan Wu, Bolin Ni, Houwen Peng, Bei Liu, Jianlong Fu, Hongyang Chao, Haibin Ling
Vision Transformer has shown great visual representation power in substantial vision tasks such as recognition and detection, and thus been attracting fast-growing efforts on manually designing more effective architectures. In this paper, we propose to use neural architecture search to automate this process, by searching not only the architecture but also the search space. The central idea is to gradually evolve different search dimensions guided by their E-T Error computed using a weight-sharing supernet. Moreover, we provide design guidelines of general vision transformers with extensive analysis according to the space searching process, which could promote the understanding of vision transformer. Remarkably, the searched models, named S3 (short for Searching the Search Space), from the searched space achieve superior performance to recently proposed models, such as Swin, DeiT and ViT, when evaluated on ImageNet. The effectiveness of S3 is also illustrated on object detection, semantic segmentation and visual question answering, demonstrating its generality to downstream vision and vision-language tasks. Code and models will be available at https://github.com/microsoft/Cream.
****
Relative stability toward diffeomorphisms indicates performance in deep nets
Leonardo Petrini, Alessandro Favero, Mario Geiger, Matthieu Wyart
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Raw Nav-merge Seismic Data to Subsurface Properties with MLP based Multi-Modal Information Unscrambler
Aditya Desai, Zhaozhuo Xu, Menal Gupta, Anu Chandran, Antoine Vial-Aussavy, Anshumali Shrivastava
Traditional seismic inversion (SI) maps the hundreds of terabytes of raw-field data to subsurface properties in gigabytes. This inversion process is expensive, requiring over a year of human and computational effort. Recently, data-driven approaches equipped with Deep learning (DL) are envisioned to improve SI efficiency. However, these improvements are restricted to data with highly reduced scale and complexity. To extend these approaches to real-scale seismic data, researchers need to process raw nav-merge seismic data into an image and perform convolution. We argue that this convolution-based way of SI is not only computationally expensive but also conceptually problematic. Seismic data is not naturally an image and need not be processed as images. In this work, we go beyond convolution and propose a novel SI method. We solve the scalability of SI by proposing a new auxiliary learning paradigm for SI (Aux-SI). This paradigm breaks the SI into local inversion tasks, which predicts each small chunk of subsurface properties using surrounding seismic data. Aux-SI combines these local predictions to obtain the entire subsurface model. However, even this local inversion is still challenging due to: (1) high-dimensional, spatially irregular multi-modal seismic data, (2) there is no concrete spatial mapping (or alignment) between subsurface properties and raw data. To handle these challenges, we propose an all-MLP architecture, Multi-Modal Information Unscrambler (MMI-Unscrambler), that unscrambles seismic information by ingesting all available multi-modal data. The experiment shows that MMI-Unscrambler outperforms both SOTA U-Net and Transformer models

on simulation data. We also scale MMI-Unscrambler to raw-field nav-merge data on Gulf-of-Mexico to obtain a geologically sound velocity model with an SSIM score of 0.8. To the best of our knowledge, this is the first successful demonstration of the DL approach on SI for real, large-scale, and complicated raw field data.
****
Inverse Problems Leveraging Pre-trained Contrastive Representations
Sriram Ravula, Georgios Smyrnis, Matt Jordan, Alexandros G. Dimakis
We study a new family of inverse problems for recovering representations of corrupted data. We assume access to a pre-trained representation learning network R(x) that operates on clean images, like CLIP. The problem is to recover the representation of an image R(x), if we are only given a corrupted version A(x), for some known forward operator A. We propose a supervised inversion method that uses a contrastive objective to obtain excellent representations for highly corrupted images. Using a linear probe on our robust representations, we achieve a higher accuracy than end-to-end supervised baselines when classifying images with various types of distortions, including blurring, additive noise, and random pixel masking. We evaluate on a subset of ImageNet and observe that our method is robust to varying levels of distortion. Our method outperforms end-to-end baselines even with a fraction of the labeled data in a wide range of forward operators.
****
The Unbalanced Gromov Wasserstein Distance: Conic Formulation and Relaxation
Thibault Sejourne, Francois-Xavier Vialard, Gabriel Peyré
Comparing metric measure spaces (i.e. a metric space endowed with a probability distribution) is at the heart of many machine learning problems. The most popular distance between such metric measure spaces is the Gromov-Wasserstein (GW) distance, which is the solution of a quadratic assignment problem. The GW distance is however limited to the comparison of metric measure spaces endowed with a \emph{probability} distribution. To alleviate this issue, we introduce two Unbalanced Gromov-Wasserstein formulations: a distance and a more tractable upper-bounding relaxation. They both allow the comparison of metric spaces equipped with arbitrary positive measures up to isometries. The first formulation is a positive and definite divergence based on a relaxation of the mass conservation constraint using a novel type of quadratically-homogeneous divergence. This divergence works hand in hand with the entropic regularization approach which is popular to solve large scale optimal transport problems. We show that the underlying non-convex optimization problem can be efficiently tackled using a highly parallelizable and GPU-friendly iterative scheme. The second formulation is a distance between mm-spaces up to isometries based on a conic lifting. Lastly, we provide numerical experiments on synthetic and domain adaptation data with a Positive-Unlabeled learning task to highlight the salient features of the unbalanced divergence and its potential applications in ML.
****
Diffusion Models Beat GANs on Image Synthesis
Prafulla Dhariwal, Alexander Nichol
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 MDPs from Features: Predict-Then-Optimize for Sequential Decision Making by Reinforcement Learning
Kai Wang, Sanket Shah, Haipeng Chen, Andrew Perrault, Finale Doshi-Velez, Milind Tambe
In the predict-then-optimize framework, the objective is to train a predictive model, mapping from environment features to parameters of an optimization problem, which maximizes decision quality when the optimization is subsequently solved. Recent work on decision-focused learning shows that embedding the optimization problem in the training pipeline can improve decision quality and help generalize better to unseen tasks compared to relying on an intermediate loss function fo

r evaluating prediction quality. We study the predict-then-optimize framework in the context of sequential decision problems (formulated as MDPs) that are solved via reinforcement learning. In particular, we are given environment features and a set of trajectories from training MDPs, which we use to train a predictive model that generalizes to unseen test MDPs without trajectories. Two significant computational challenges arise in applying decision-focused learning to MDPs: (i) large state and action spaces make it infeasible for existing techniques to differentiate through MDP problems, and (ii) the high-dimensional policy space, as parameterized by a neural network, makes differentiating through a policy expensive. We resolve the first challenge by sampling provably unbiased derivatives to approximate and differentiate through optimality conditions, and the second challenge by using a low-rank approximation to the high-dimensional sample-based derivatives. We implement both Bellman-based and policy gradient-based decision-focused learning on three different MDP problems with missing parameters, and show that decision-focused learning performs better in generalization to unseen tasks.
****
A Closer Look at the Worst-case Behavior of Multi-armed Bandit Algorithms
Anand Kalvit, Assaf Zeevi
One of the key drivers of complexity in the classical (stochastic) multi-armed bandit (MAB) problem is the difference between mean rewards in the top two arms, also known as the instance gap. The celebrated Upper Confidence Bound (UCB) policy is among the simplest optimism-based MAB algorithms that naturally adapts to this gap: for a horizon of play n, it achieves optimal O(log n) regret in instances with "large" gaps, and a near-optimal O(\sqrt{n log n}) minimax regret when the gap can be arbitrarily "small." This paper provides new results on the arm-sampling behavior of UCB, leading to several important insights. Among these, it is shown that arm-sampling rates under UCB are asymptotically deterministic, regardless of the problem complexity. This discovery facilitates new sharp asymptotics and a novel alternative proof for the O(\sqrt{n log n}) minimax regret of UCB. Furthermore, the paper also provides the first complete process-level characterization of the MAB problem in the conventional diffusion scaling. Among other things, the "small" gap worst-case lens adopted in this paper also reveals profound distinctions between the behavior of UCB and Thompson Sampling, such as an "incomplete learning" phenomenon characteristic of the latter.
****
SAPE: Spatially-Adaptive Progressive Encoding for Neural Optimization
Amir Hertz, Or Perel, Raja Giryes, Olga Sorkine-hornung, Daniel Cohen-or
Multilayer-perceptrons (MLP) are known to struggle learning functions of high-frequencies, and in particular, instances of wide frequency bands.We present a progressive mapping scheme for input signals of MLP networks, enabling them to better fit a wide range of frequencies without sacrificing training stability or requiring any domain specific preprocessing. We introduce Spatially Adaptive Progressive Encoding (SAPE) layers, which gradually unmask signal components with increasing frequencies as a function of time and space. The progressive exposure of frequencies is monitored by a feedback loop throughout the neural optimization process, allowing changes to propagate at different rates among local spatial portions of the signal space. We demonstrate the advantage of our method on variety of domains and applications: regression of low dimensional signals and images, representation learning of occupancy networks, and a geometric task of mesh transfer between 3D shapes.
****
A Biased Graph Neural Network Sampler with Near-Optimal Regret
Qingru Zhang, David Wipf, Quan Gan, Le Song
Graph neural networks (GNN) have recently emerged as a vehicle for applying deep network architectures to graph and relational data. However, given the increasing size of industrial datasets, in many practical situations, the message passing computations required for sharing information across GNN layers are no longer scalable. Although various sampling methods have been introduced to approximate full-graph training within a tractable budget, there remain unresolved complica

tions such as high variances and limited theoretical guarantees. To address the se issues, we build upon existing work and treat GNN neighbor sampling as a mult i-armed bandit problem but with a newly-designed reward function that introduces some degree of bias designed to reduce variance and avoid unstable, possibly-un bounded pay outs. And unlike prior bandit-GNN use cases, the resulting policy l eads to near-optimal regret while accounting for the GNN training dynamics intro duced by SGD. From a practical standpoint, this translates into lower variance e stimates and competitive or superior test accuracy across several benchmarks.
****
Equilibrium Refinement for the Age of Machines: The One-Sided Quasi-Perfect Equi librium
Gabriele Farina, Tuomas Sandholm
In two-player zero-sum extensive-form games, Nash equilibrium prescribes optimal strategies against perfectly rational opponents. However, it does not guarantee rational play in parts of the game tree that can only be reached by the players making mistakes. This can be problematic when operationalizing equilibria in th e real world among imperfect players. Trembling-hand refinements are a sound rem edy to this issue, and are subsets of Nash equilibria that are designed to handl e the possibility that any of the players may make mistakes. In this paper, we i nitiate the study of equilibrium refinements for settings where one of the playe rs is perfectly rational (the ``machine'') and the other may make mistakes. As w e show, this endeavor has many pitfalls: many intuitively appealing approaches t o refinement fail in various ways. On the positive side, we introduce a modifica tion of the classical quasi-perfect equilibrium (QPE) refinement, which we call the one-sided quasi-perfect equilibrium. Unlike QPE, one-sided QPE only accounts for mistakes from one player and assumes that no mistakes will be made by the m achine. We present experiments on standard benchmark games and an endgame from t he famous man-machine match where the AI Libratus was the first to beat top huma n specialist professionals in heads-up no-limit Texas hold'em poker. We show tha t one-sided QPE can be computed more efficiently than all known prior refinement s, paving the way to wider adoption of Nash equilibrium refinements in settings with perfectly rational machines (or humans perfectly actuating machine-generate d strategies) that interact with players prone to mistakes. We also show that on e-sided QPE tends to play better than a Nash equilibrium strategy against imperf ect opponents.
****
Interpreting Representation Quality of DNNs for 3D Point Cloud Processing
Wen Shen, Qihan Ren, Dongrui Liu, Quanshi Zhang
In this paper, we evaluate the quality of knowledge representations encoded in d eep neural networks (DNNs) for 3D point cloud processing. We propose a method to disentangle the overall model vulnerability into the sensitivity to the rotatio n, the translation, the scale, and local 3D structures. Besides, we also propose metrics to evaluate the spatial smoothness of encoding 3D structures, and the r epresentation complexity of the DNN. Based on such analysis, experiments expose representation problems with classic DNNs, and explain the utility of the advers arial training. The code will be released when this paper is accepted.
****
How Fine-Tuning Allows for Effective Meta-Learning
Kurtland Chua, Qi Lei, Jason D. Lee
Representation learning has served as a key tool for meta-learning, enabling rap id learning of new tasks. Recent works like MAML learn task-specific representat ions by finding an initial representation requiring minimal per-task adaptation (i.e. a fine-tuning-based objective). We present a theoretical framework for ana lyzing a MAML-like algorithm, assuming all available tasks require approximately the same representation. We then provide risk bounds on predictors found by fin e-tuning via gradient descent, demonstrating that the method provably leverages the shared structure. We illustrate these bounds in the logistic regression and neural network settings. In contrast, we establish settings where learning one r epresentation for all tasks (i.e. using a "frozen representation" objective) fai ls. Notably, any such algorithm cannot outperform directly learning the target t

ask with no other information, in the worst case. This separation underscores the benefit of fine-tuning-based over “frozen representation” objectives in few-shot learning.
****
Cooperative Stochastic Bandits with Asynchronous Agents and Constrained Feedback
Lin Yang, Yu-Zhen Janice Chen, Stephen Pasteris, Mohammad Hajiesmaili, John C. S. Lui, Don Towsley
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Multiple Descent: Design Your Own Generalization Curve
Lin Chen, Yifei Min, Mikhail Belkin, Amin Karbasi
This paper explores the generalization loss of linear regression in variably parameterized families of models, both under-parameterized and over-parameterized. We show that the generalization curve can have an arbitrary number of peaks, and moreover, the locations of those peaks can be explicitly controlled. Our results highlight the fact that both the classical U-shaped generalization curve and the recently observed double descent curve are not intrinsic properties of the model family. Instead, their emergence is due to the interaction between the properties of the data and the inductive biases of learning algorithms.
****
On Empirical Risk Minimization with Dependent and Heavy-Tailed Data
Abhishek Roy, Krishnakumar Balasubramanian, Murat A. Erdogdu
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Gone Fishing: Neural Active Learning with Fisher Embeddings
Jordan Ash, Surbhi Goel, Akshay Krishnamurthy, Sham Kakade
There is an increasing need for effective active learning algorithms that are compatible with deep neural networks. This paper motivates and revisits a classic, Fisher-based active selection objective, and proposes BAIT, a practical, tractable, and high-performing algorithm that makes it viable for use with neural models. BAIT draws inspiration from the theoretical analysis of maximum likelihood estimators (MLE) for parametric models. It selects batches of samples by optimizing a bound on the MLE error in terms of the Fisher information, which we show can be implemented efficiently at scale by exploiting linear-algebraic structure especially amenable to execution on modern hardware. Our experiments demonstrate that BAIT outperforms the previous state of the art on both classification and regression problems, and is flexible enough to be used with a variety of model architectures.
****
On Riemannian Optimization over Positive Definite Matrices with the Bures-Wasserstein Geometry
Andi Han, Bamdev Mishra, Pratik Kumar Jawanpuria, Junbin Gao
In this paper, we comparatively analyze the Bures-Wasserstein (BW) geometry with the popular Affine-Invariant (AI) geometry for Riemannian optimization on the symmetric positive definite (SPD) matrix manifold. Our study begins with an observation that the BW metric has a linear dependence on SPD matrices in contrast to the quadratic dependence of the AI metric. We build on this to show that the BW metric is a more suitable and robust choice for several Riemannian optimization problems over ill-conditioned SPD matrices. We show that the BW geometry has a non-negative curvature, which further improves convergence rates of algorithms over the non-positively curved AI geometry. Finally, we verify that several popular cost functions, which are known to be geodesic convex under the AI geometry, are also geodesic convex under the BW geometry. Extensive experiments on various applications support our findings.

****
Refining Language Models with Compositional Explanations
Huihan Yao, Ying Chen, Qinyuan Ye, Xisen Jin, Xiang Ren
Pre-trained language models have been successful on text classification tasks, but are prone to learning spurious correlations from biased datasets, and are thus vulnerable when making inferences in a new domain. Prior work reveals such spurious patterns via post-hoc explanation algorithms which compute the importance of input features. Further, the model is regularized to align the importance scores with human knowledge, so that the unintended model behaviors are eliminated. However, such a regularization technique lacks flexibility and coverage, since only importance scores towards a pre-defined list of features are adjusted, while more complex human knowledge such as feature interaction and pattern generalization can hardly be incorporated. In this work, we propose to refine a learned language model for a target domain by collecting human-provided compositional explanations regarding observed biases. By parsing these explanations into executable logic rules, the human-specified refinement advice from a small set of explanations can be generalized to more training examples. We additionally introduce a regularization term allowing adjustments for both importance and interaction of features to better rectify model behavior. We demonstrate the effectiveness of the proposed approach on two text classification tasks by showing improved performance in target domain as well as improved model fairness after refinement.
****
Going Beyond Linear RL: Sample Efficient Neural Function Approximation
Baihe Huang, Kaixuan Huang, Sham Kakade, Jason D. Lee, Qi Lei, Runzhe Wang, Jiaqi Yang
Deep Reinforcement Learning (RL) powered by neural net approximation of the Q function has had enormous empirical success. While the theory of RL has traditionally focused on linear function approximation (or eluder dimension) approaches, little is known about nonlinear RL with neural net approximations of the Q functions. This is the focus of this work, where we study function approximation with two-layer neural networks (considering both ReLU and polynomial activation functions). Our first result is a computationally and statistically efficient algorithm in the generative model setting under completeness for two-layer neural networks. Our second result considers this setting but under only realizability of the neural net function class. Here, assuming deterministic dynamics, the sample complexity scales linearly in the algebraic dimension. In all cases, our results significantly improve upon what can be attained with linear (or eluder dimension) methods.
****
Scalable Neural Data Server: A Data Recommender for Transfer Learning
Tianshi Cao, Sasha (Alexandre) Doubov, David Acuna, Sanja Fidler
Absence of large-scale labeled data in the practitioner's target domain can be a bottleneck to applying machine learning algorithms in practice. Transfer learning is a popular strategy for leveraging additional data to improve the downstream performance, but finding the most relevant data to transfer from can be challenging. Neural Data Server (NDS), a search engine that recommends relevant data for a given downstream task, has been previously proposed to address this problem (Yan et al., 2020). NDS uses a mixture of experts trained on data sources to estimate similarity between each source and the downstream task. Thus, the computational cost to each user grows with the number of sources and requires an expensive training step for each data provider.To address these issues, we propose Scalable Neural Data Server (SNDS), a large-scale search engine that can theoretically index thousands of datasets to serve relevant ML data to end users. SNDS trains the mixture of experts on intermediary datasets during initialization, and represents both data sources and downstream tasks by their proximity to the intermediary datasets. As such, computational cost incurred by users of SNDS remains fixed as new datasets are added to the server, without pre-training for the data providers.We validate SNDS on a plethora of real world tasks and find that data recommended by SNDS improves downstream task performance over baselines. We also demonstrate the scalability of our system by demonstrating its ability to sele

ct relevant data for transfer outside of the natural image setting.
****
What can linearized neural networks actually say about generalization?
Guillermo Ortiz-Jimenez, Seyed-Mohsen Moosavi-Dezfooli, Pascal Frossard
For certain infinitely-wide neural networks, the neural tangent kernel (NTK) theory fully characterizes generalization, but for the networks used in practice, the empirical NTK only provides a rough first-order approximation. Still, a growing body of work keeps leveraging this approximation to successfully analyze important deep learning phenomena and design algorithms for new applications. In our work, we provide strong empirical evidence to determine the practical validity of such approximation by conducting a systematic comparison of the behavior of different neural networks and their linear approximations on different tasks. We show that the linear approximations can indeed rank the learning complexity of certain tasks for neural networks, even when they achieve very different performances. However, in contrast to what was previously reported, we discover that neural networks do not always perform better than their kernel approximations, and reveal that the performance gap heavily depends on architecture, dataset size and training task. We discover that networks overfit to these tasks mostly due to the evolution of their kernel during training, thus, revealing a new type of implicit bias.
****
CATs: Cost Aggregation Transformers for Visual Correspondence
Seokju Cho, Sunghwan Hong, Sangryul Jeon, Yunsung Lee, Kwanghoon Sohn, Seungryong Kim
We propose a novel cost aggregation network, called Cost Aggregation Transformers (CATs), to find dense correspondences between semantically similar images with additional challenges posed by large intra-class appearance and geometric variations. Cost aggregation is a highly important process in matching tasks, which the matching accuracy depends on the quality of its output. Compared to hand-crafted or CNN-based methods addressing the cost aggregation, in that either lacks robustness to severe deformations or inherit the limitation of CNNs that fail to discriminate incorrect matches due to limited receptive fields, CATs explore global consensus among initial correlation map with the help of some architectural designs that allow us to fully leverage self-attention mechanism. Specifically, we include appearance affinity modeling to aid the cost aggregation process in order to disambiguate the noisy initial correlation maps and propose multi-level aggregation to efficiently capture different semantics from hierarchical feature representations. We then combine with swapping self-attention technique and residual connections not only to enforce consistent matching, but also to ease the learning process, which we find that these result in an apparent performance boost. We conduct experiments to demonstrate the effectiveness of the proposed model over the latest methods and provide extensive ablation studies. Code and trained models are available at https://sunghwanhong.github.io/CATs/.
****
Asynchronous Stochastic Optimization Robust to Arbitrary Delays
Alon Cohen, Amit Daniely, Yoel Drori, Tomer Koren, Mariano Schain
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Consistent Non-Parametric Methods for Maximizing Robustness
Robi Bhattacharjee, Kamalika Chaudhuri
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Generalizable Multi-linear Attention Network
Tao Jin, Zhou Zhao

The majority of existing multimodal sequential learning methods focus on how to obtain effective representations and ignore the importance of multimodal fusion. Bilinear attention network (BAN) is a commonly used fusion method, which leverages tensor operations to associate the features of different modalities. However, BAN has a poor compatibility for more modalities, since the computational complexity of the attention map increases exponentially with the number of modalities. Based on this concern, we propose a new method called generalizable multi-linear attention network (MAN), which can associate as many modalities as possible in linear complexity with hierarchical approximation decomposition (HAD). Besides, considering the fact that softmax attention kernels cannot be decomposed as linear operation directly, we adopt the addition random features (ARF) mechanism to approximate the non-linear softmax functions with enough theoretical analysis. We conduct extensive experiments on four datasets of three tasks (multimodal sentiment analysis, multimodal speaker traits recognition, and video retrieval), the experimental results show that MAN could achieve competitive results compared with the state-of-the-art methods, showcasing the effectiveness of the approximation decomposition and addition random features mechanism.
****
Labeling Trick: A Theory of Using Graph Neural Networks for Multi-Node Representation Learning
Muhan Zhang, Pan Li, Yinglong Xia, Kai Wang, Long Jin
In this paper, we provide a theory of using graph neural networks (GNNs) for multi-node representation learning (where we are interested in learning a representation for a set of more than one node, such as link). We know that GNN is designed to learn single-node representations. When we want to learn a node set representation involving multiple nodes, a common practice in previous works is to directly aggregate the single-node representations obtained by a GNN into a joint node set representation. In this paper, we show a fundamental constraint of such an approach, namely the inability to capture the dependence between nodes in the node set, and argue that directly aggregating individual node representations does not lead to an effective joint representation for multiple nodes. Then, we notice that a few previous successful works for multi-node representation learning, including SEAL, Distance Encoding, and ID-GNN, all used node labeling. These methods first label nodes in the graph according to their relationships with the target node set before applying a GNN. Then, the node representations obtained in the labeled graph are aggregated into a node set representation. By investigating their inner mechanisms, we unify these node labeling techniques into a single and most general form---labeling trick. We prove that with labeling trick a sufficiently expressive GNN learns the most expressive node set representations, thus in principle solves any joint learning tasks over node sets. Experiments on one important two-node representation learning task, link prediction, verified our theory. Our work explains the superior performance of previous node-labeling-based methods, and establishes a theoretical foundation of using GNNs for multi-node representation learning.
****
SUPER-ADAM: Faster and Universal Framework of Adaptive Gradients
Feihu Huang, Junyi Li, Heng Huang
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
General Nonlinearities in SO(2)-Equivariant CNNs
Daniel Franzen, Michael Wand
Invariance under symmetry is an important problem in machine learning. Our paper looks specifically at equivariant neural networks where transformations of inputs yield homomorphic transformations of outputs. Here, steerable CNNs have emerged as the standard solution. An inherent problem of steerable representations is that general nonlinear layers break equivariance, thus restricting architectural choices. Our paper applies harmonic distortion analysis to illuminate the effe

ct of nonlinearities on Fourier representations of SO(2). We develop a novel FFT -based algorithm for computing representations of non-linearly transformed activ ations while maintaining band-limitation. It yields exact equivariance for polyn omial (approximations of) nonlinearities, as well as approximate solutions with tunable accuracy for general functions. We apply the approach to build a fully E (3)-equivariant network for sampled 3D surface data. In experiments with 2D and 3D data, we obtain results that compare favorably to the state-of-the-art in ter ms of accuracy while permitting continuous symmetry and exact equivariance.
****
Denoising Normalizing Flow
Christian Horvat, Jean-Pascal Pfister
Normalizing flows (NF) are expressive as well as tractable density estimation me thods whenever the support of the density is diffeomorphic to the entire data-sp ace. However, real-world data sets typically live on (or very close to) low-dime nsional manifolds thereby challenging the applicability of standard NF on real-w orld problems. Here we propose a novel method - called Denoising Normalizing Flo w (DNF) - that estimates the density on the low-dimensional manifold while learn ing the manifold as well. The DNF works in 3 steps. First, it inflates the manif old - making it diffeomorphic to the entire data-space. Secondly, it learns an N F on the inflated manifold and finally it learns a denoising mapping - similarly to denoising autoencoders. The DNF relies on a single cost function and does no t require to alternate between a density estimation phase and a manifold learnin g phase - as it is the case with other recent methods. Furthermore, we show that the DNF can learn meaningful low-dimensional representations from naturalistic images as well as generate high-quality samples.
****
Attention over Learned Object Embeddings Enables Complex Visual Reasoning
David Ding, Felix Hill, Adam Santoro, Malcolm Reynolds, Matt Botvinick
Neural networks have achieved success in a wide array of perceptual tasks but of ten fail at tasks involving both perception and higher-level reasoning. On these more challenging tasks, bespoke approaches (such as modular symbolic components , independent dynamics models or semantic parsers) targeted towards that specifi c type of task have typically performed better. The downside to these targeted a pproaches, however, is that they can be more brittle than general-purpose neural networks, requiring significant modification or even redesign according to the particular task at hand. Here, we propose a more general neural-network-based ap proach to dynamic visual reasoning problems that obtains state-of-the-art perfor mance on three different domains, in each case outperforming bespoke modular app roaches tailored specifically to the task. Our method relies on learned object-c entric representations, self-attention and self-supervised dynamics learning, an d all three elements together are required for strong performance to emerge. The success of this combination suggests that there may be no need to trade off fle xibility for performance on problems involving spatio-temporal or causal-style r easoning. With the right soft biases and learning objectives in a neural network we may be able to attain the best of both worlds.
****
Differentially Private Federated Bayesian Optimization with Distributed Explorat ion
Zhongxiang Dai, Bryan Kian Hsiang Low, Patrick Jaillet
Bayesian optimization (BO) has recently been extended to the federated learning (FL) setting by the federated Thompson sampling (FTS) algorithm, which has promi sing applications such as federated hyperparameter tuning. However, FTS is not e quipped with a rigorous privacy guarantee which is an important consideration in FL. Recent works have incorporated differential privacy (DP) into the training of deep neural networks through a general framework for adding DP to iterative a lgorithms. Following this general DP framework, our work here integrates DP into FTS to preserve user-level privacy. We also leverage the ability of this genera l DP framework to handle different parameter vectors, as well as the technique o f local modeling for BO, to further improve the utility of our algorithm through distributed exploration (DE). The resulting differentially private FTS with DE

(DP-FTS-DE) algorithm is endowed with theoretical guarantees for both the privacy and utility and is amenable to interesting theoretical insights about the privacy-utility trade-off. We also use real-world experiments to show that DP-FTS-DE achieves high utility (competitive performance) with a strong privacy guarantee (small privacy loss) and induces a trade-off between privacy and utility.
****
Differentiable Learning Under Triage
Nastaran Okati, Abir De, Manuel Rodriguez
Multiple lines of evidence suggest that predictive models may benefit from algorithmic triage. Under algorithmic triage, a predictive model does not predict all instances but instead defers some of them to human experts. However, the interplay between the prediction accuracy of the model and the human experts under algorithmic triage is not well understood. In this work, we start by formally characterizing under which circumstances a predictive model may benefit from algorithmic triage. In doing so, we also demonstrate that models trained for full automation may be suboptimal under triage. Then, given any model and desired level of triage, we show that the optimal triage policy is a deterministic threshold rule in which triage decisions are derived deterministically by thresholding the difference between the model and human errors on a per-instance level. Building upon these results, we introduce a practical gradient-based algorithm that is guaranteed to find a sequence of predictive models and triage policies of increasing performance. Experiments on a wide variety of supervised learning tasks using synthetic and real data from two important applications---content moderation and scientific discovery---illustrate our theoretical results and show that the models and triage policies provided by our gradient-based algorithm outperform those provided by several competitive baselines.
****
ROI Maximization in Stochastic Online Decision-Making
Nicolò Cesa-Bianchi, Tom Cesari, Yishay Mansour, 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.
****
When Expressivity Meets Trainability: Fewer than n Neurons Can Work
Jiawei Zhang, Yushun Zhang, Mingyi Hong, Ruoyu Sun, Zhi-Quan 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.
****
Analyzing the Confidentiality of Undistillable Teachers in Knowledge Distillation
Souvik Kundu, Qirui Sun, Yao Fu, Massoud Pedram, Peter Beerel
Knowledge distillation (KD) has recently been identified as a method that can unintentionally leak private information regarding the details of a teacher model to an unauthorized student. Recent research in developing undistillable nasty teachers that can protect model confidentiality has gained significant attention. However, the level of protection these nasty models offer has been largely untested. In this paper, we show that transferring knowledge to a shallow sub-section of a student can largely reduce a teacher’s influence. By exploring the depth of the shallow subsection, we then present a distillation technique that enables a skeptical student model to learn even from a nasty teacher. To evaluate the efficacy of our skeptical students, we conducted experiments with several models with KD on both training data-available and data-free scenarios for various datasets. While distilling from nasty teachers, compared to the normal student models, skeptical students consistently provide superior classification performance of up to ~59.5%. Moreover, similar to normal students, skeptical students maintain high classification accuracy when distilled from a normal teacher, showing their efficacy irrespective of the teacher being nasty or not. We believe the abilit

y of skeptical students to largely diminish the KD-immunity of potentially nasty teachers will motivate the research community to create more robust mechanisms for model confidentiality. We have open-sourced the code at https://github.com/ksouvik52/Skeptical2021
****
High Probability Complexity Bounds for Line Search Based on Stochastic Oracles
Billy Jin, Katya Scheinberg, Miaolan Xie
We consider a line-search method for continuous optimization under a stochastic setting where the function values and gradients are available only through inexact probabilistic zeroth and first-order oracles. These oracles capture multiple standard settings including expected loss minimization and zeroth-order optimization. Moreover, our framework is very general and allows the function and gradient estimates to be biased. The proposed algorithm is simple to describe, easy to implement, and uses these oracles in a similar way as the standard deterministic line search uses exact function and gradient values. Under fairly general conditions on the oracles, we derive a high probability tail bound on the iteration complexity of the algorithm when applied to non-convex smooth functions. These results are stronger than those for other existing stochastic line search methods and apply in more general settings.
****
Pay Attention to MLPs
Hanxiao Liu, Zihang Dai, David So, Quoc V Le
Transformers have become one of the most important architectural innovations in deep learning and have enabled many breakthroughs over the past few years. Here we propose a simple network architecture, gMLP, based solely on MLPs with gating, and show that it can perform as well as Transformers in key language and vision applications. Our comparisons show that self-attention is not critical for Vision Transformers, as gMLP can achieve the same accuracy. For BERT, our model achieves parity with Transformers on pretraining perplexity and is better on some downstream NLP tasks. On finetuning tasks where gMLP performs worse, making the gMLP model substantially larger can close the gap with Transformers. In general, our experiments show that gMLP can scale as well as Transformers over increased data and compute.
****
An Image is Worth More Than a Thousand Words: Towards Disentanglement in The Wild
Aviv Gabbay, Niv Cohen, Yedid Hoshen
Unsupervised disentanglement has been shown to be theoretically impossible without inductive biases on the models and the data. As an alternative approach, recent methods rely on limited supervision to disentangle the factors of variation and allow their identifiability. While annotating the true generative factors is only required for a limited number of observations, we argue that it is infeasible to enumerate all the factors of variation that describe a real-world image distribution. To this end, we propose a method for disentangling a set of factors which are only partially labeled, as well as separating the complementary set of residual factors that are never explicitly specified. Our success in this challenging setting, demonstrated on synthetic benchmarks, gives rise to leveraging off-the-shelf image descriptors to partially annotate a subset of attributes in real image domains (e.g. of human faces) with minimal manual effort. Specifically, we use a recent language-image embedding model (CLIP) to annotate a set of attributes of interest in a zero-shot manner and demonstrate state-of-the-art disentangled image manipulation results.
****
Dynamics of Stochastic Momentum Methods on Large-scale, Quadratic Models
Courtney Paquette, Elliot Paquette
We analyze a class of stochastic gradient algorithms with momentum on a high-dimensional random least squares problem. Our framework, inspired by random matrix theory, provides an exact (deterministic) characterization for the sequence of function values produced by these algorithms which is expressed only in terms of the eigenvalues of the Hessian. This leads to simple expressions for nearly-opti

mal hyperparameters, a description of the limiting neighborhood, and average-cas
e complexity. As a consequence, we show that (small-batch) stochastic heavy-ball
 momentum with a fixed momentum parameter provides no actual performance improve
ment over SGD when step sizes are adjusted correctly. For contrast, in the non-s
trongly convex setting, it is possible to get a large improvement over SGD using
 momentum. By introducing hyperparameters that depend on the number of samples,
we propose a new algorithm sDANA (stochastic dimension adjusted Nesterov acceler
ation) which obtains an asymptotically optimal average-case complexity while rem
aining linearly convergent in the strongly convex setting without adjusting para
meters.
****
Adversarial Examples in Multi-Layer Random ReLU Networks
Peter Bartlett, Sebastien Bubeck, Yeshwanth Cherapanamjeri
We consider the phenomenon of adversarial examples in ReLU networks with indepen
dent Gaussian parameters. For networks of constant depth and with a large range
 of widths (for instance, it suffices if the width of each layer is polynomial i
n that of any other layer), small perturbations of input vectors lead to large c
hanges of outputs. This generalizes results of Daniely and Schacham (2020) for
networks of rapidly decreasing width and of Bubeck et al (2021) for two-layer ne
tworks. Our proof shows that adversarial examples arise in these networks becaus
e the functions they compute are \emph{locally} very similar to random linear fu
nctions. Bottleneck layers play a key role: the minimal width up to some point i
n the network determines scales and sensitivities of mappings computed up to tha
t point. The main result is for networks with constant depth, but we also show
that some constraint on depth is necessary for a result of this kind, because th
ere are suitably deep networks that, with constant probability, compute a functi
on that is close to constant.
****
Efficient Statistical Assessment of Neural Network Corruption Robustness
Karim TIT, Teddy Furon, Mathias ROUSSET
We quantify the robustness of a trained network to input uncertainties with a st
ochastic simulation inspired by the field of Statistical Reliability Engineering
. The robustness assessment is cast as a statistical hypothesis test: the networ
k is deemed as locally robust if the estimated probability of failure is lower t
han a critical level.The procedure is based on an Importance Splitting simulatio
n generating samples of rare events. We derive theoretical guarantees that are n
on-asymptotic w.r.t. sample size. Experiments tackling large scale networks outl
ine the efficiency of our method making a low number of calls to the network fun
ction.
****
A Highly-Efficient Group Elastic Net Algorithm with an Application to Function-O
n-Scalar Regression
Tobia Boschi, Matthew Reimherr, Francesca Chiaromonte
Feature Selection and Functional Data Analysis are two dynamic areas of research
, with important applications in the analysis of large and complex data sets. St
raddling these two areas, we propose a new highly efficient algorithm to perform
 Group Elastic Net with application to function-on-scalar feature selection, whe
re a functional response is modeled against a very large number of potential sca
lar predictors. First, we introduce a new algorithm to solve Group Elastic Net i
n ultra-high dimensional settings, which exploits the sparsity structure of the
Augmented Lagrangian to greatly reduce the computational burden. Next, taking ad
vantage of the properties of Functional Principal Components, we extend our algo
rithm to the function-on-scalar regression framework. We use simulations to demo
nstrate the CPU time gains afforded by our approach compared to its best existin
g competitors, and present an application to data from a Genome-Wide Association
 Study on childhood obesity.
****
Hierarchical Clustering: $O(1)$-Approximation for Well-Clustered Graphs
Bogdan-Adrian Manghiuc, He 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.
****
Realistic evaluation of transductive few-shot learning
Olivier Veilleux, Malik Boudiaf, Pablo Piantanida, Ismail Ben Ayed
Transductive inference is widely used in few-shot learning, as it leverages the statistics of the unlabeled query set of a few-shot task, typically yielding substantially better performances than its inductive counterpart. The current few-shot benchmarks use perfectly class-balanced tasks at inference. We argue that such an artificial regularity is unrealistic, as it assumes that the marginal label probability of the testing samples is known and fixed to the uniform distribution. In fact, in realistic scenarios, the unlabeled query sets come with arbitrary and unknown label marginals. We introduce and study the effect of arbitrary class distributions within the query sets of few-shot tasks at inference, removing the class-balance artefact. Specifically, we model the marginal probabilities of the classes as Dirichlet-distributed random variables, which yields a principled and realistic sampling within the simplex. This leverages the current few-shot benchmarks, building testing tasks with arbitrary class distributions. We evaluate experimentally state-of-the-art transductive methods over 3 widely used data sets, and observe, surprisingly, substantial performance drops, even below inductive methods in some cases. Furthermore, we propose a generalization of the mutual-information loss, based on α-divergences, which can handle effectively class-distribution variations. Empirically, we show that our transductive α-divergence optimization outperforms state-of-the-art methods across several data sets, models and few-shot settings.
****
Qu-ANTI-zation: Exploiting Quantization Artifacts for Achieving Adversarial Outcomes
Sanghyun Hong, Michael-Andrei Panaitescu-Liess, Yigitcan Kaya, Tudor Dumitras
Quantization is a popular technique that transforms the parameter representation of a neural network from floating-point numbers into lower-precision ones (e.g., 8-bit integers). It reduces the memory footprint and the computational cost at inference, facilitating the deployment of resource-hungry models. However, the parameter perturbations caused by this transformation result in behavioral disparities between the model before and after quantization. For example, a quantized model can misclassify some test-time samples that are otherwise classified correctly. It is not known whether such differences lead to a new security vulnerability. We hypothesize that an adversary may control this disparity to introduce specific behaviors that activate upon quantization. To study this hypothesis, we weaponize quantization-aware training and propose a new training framework to implement adversarial quantization outcomes. Following this framework, we present three attacks we carry out with quantization: (i) an indiscriminate attack for significant accuracy loss; (ii) a targeted attack against specific samples; and (iii) a backdoor attack for controlling the model with an input trigger. We further show that a single compromised model defeats multiple quantization schemes, including robust quantization techniques. Moreover, in a federated learning scenario, we demonstrate that a set of malicious participants who conspire can inject our quantization-activated backdoor. Lastly, we discuss potential counter-measures and show that only re-training consistently removes the attack artifacts. Our code is available at https://github.com/Secure-AI-Systems-Group/Qu-ANTI-zation
****
Differentially Private Stochastic Optimization: New Results in Convex and Non-Convex Settings
Raef Bassily, Cristóbal Guzmán, Michael Menart
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

TacticZero: Learning to Prove Theorems from Scratch with Deep Reinforcement Learning
Minchao Wu, Michael Norrish, Christian Walder, Amir Dezfouli
We propose a novel approach to interactive theorem-proving (ITP) using deep reinforcement learning. The proposed framework is able to learn proof search strategies as well as tactic and arguments prediction in an end-to-end manner. We formulate the process of ITP as a Markov decision process (MDP) in which each state represents a set of potential derivation paths. This structure allows us to introduce a novel backtracking mechanism which enables the agent to efficiently discard (predicted) dead-end derivations and restart the derivation from promising alternatives. We implement the framework in the HOL theorem prover. Experimental results show that the framework using learned search strategies outperforms existing automated theorem provers (i.e., hammers) available in HOL when evaluated on unseen problems. We further elaborate the role of key components of the framework using ablation studies.
****
Integrating Tree Path in Transformer for Code Representation
Han Peng, Ge Li, Wenhan Wang, YunFei Zhao, Zhi Jin
Learning distributed representation of source code requires modelling its syntax and semantics. Recent state-of-the-art models leverage highly structured source code representations, such as the syntax trees and paths therein. In this paper, we investigate two representative path encoding methods shown in previous research work and integrate them into the attention module of Transformer. We draw inspiration from the ideas of positional encoding and modify them to incorporate these path encoding. Specifically, we encode both the pairwise path between tokens of source code and the path from the leaf node to the tree root for each token in the syntax tree. We explore the interaction between these two kinds of paths by integrating them into the unified Transformer framework. The detailed empirical study for path encoding methods also leads to our novel state-of-the-art representation model TPTrans, which finally outperforms strong baselines. Extensive experiments and ablation studies on code summarization across four different languages demonstrate the effectiveness of our approaches. We release our code at \url{https://github.com/AwdHanPeng/TPTrans}.
****
Twins: Revisiting the Design of Spatial Attention in Vision Transformers
Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, Chunhua Shen
Very recently, a variety of vision transformer architectures for dense prediction tasks have been proposed and they show that the design of spatial attention is critical to their success in these tasks. In this work, we revisit the design of the spatial attention and demonstrate that a carefully devised yet simple spatial attention mechanism performs favorably against the state-of-the-art schemes. As a result, we propose two vision transformer architectures, namely, Twins- PCPVT and Twins-SVT. Our proposed architectures are highly efficient and easy to implement, only involving matrix multiplications that are highly optimized in modern deep learning frameworks. More importantly, the proposed architectures achieve excellent performance on a wide range of visual tasks including image-level classification as well as dense detection and segmentation. The simplicity and strong performance suggest that our proposed architectures may serve as stronger backbones for many vision tasks.
****
Evaluating State-of-the-Art Classification Models Against Bayes Optimality
Ryan Theisen, Huan Wang, Lav R. Varshney, Caiming Xiong, Richard Socher
Evaluating the inherent difficulty of a given data-driven classification problem is important for establishing absolute benchmarks and evaluating progress in the field. To this end, a natural quantity to consider is the \emph{Bayes error}, which measures the optimal classification error theoretically achievable for a given data distribution. While generally an intractable quantity, we show that we can compute the exact Bayes error of generative models learned using normalizing flows. Our technique relies on a fundamental result, which states that the Ba

yes error is invariant under invertible transformation. Therefore, we can compute the exact Bayes error of the learned flow models by computing it for Gaussian base distributions, which can be done efficiently using Holmes-Diaconis-Ross integration. Moreover, we show that by varying the temperature of the learned flow models, we can generate synthetic datasets that closely resemble standard benchmark datasets, but with almost any desired Bayes error. We use our approach to conduct a thorough investigation of state-of-the-art classification models, and find that in some --- but not all --- cases, these models are capable of obtaining accuracy very near optimal. Finally, we use our method to evaluate the intrinsic "hardness" of standard benchmark datasets.
****

Data-Efficient Instance Generation from Instance Discrimination
Ceyuan Yang, Yujun Shen, Yinghao Xu, Bolei 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.
****

Reliable Post hoc Explanations: Modeling Uncertainty in Explainability
Dylan Slack, Anna Hilgard, Sameer Singh, Himabindu Lakkaraju
As black box explanations are increasingly being employed to establish model credibility in high stakes settings, it is important to ensure that these explanations are accurate and reliable. However, prior work demonstrates that explanations generated by state-of-the-art techniques are inconsistent, unstable, and provide very little insight into their correctness and reliability. In addition, these methods are also computationally inefficient, and require significant hyper-parameter tuning. In this paper, we address the aforementioned challenges by developing a novel Bayesian framework for generating local explanations along with their associated uncertainty. We instantiate this framework to obtain Bayesian versions of LIME and KernelSHAP which output credible intervals for the feature importances, capturing the associated uncertainty. The resulting explanations not only enable us to make concrete inferences about their quality (e.g., there is a 95% chance that the feature importance lies within the given range), but are also highly consistent and stable. We carry out a detailed theoretical analysis that leverages the aforementioned uncertainty to estimate how many perturbations to sample, and how to sample for faster convergence. This work makes the first attempt at addressing several critical issues with popular explanation methods in one shot, thereby generating consistent, stable, and reliable explanations with guarantees in a computationally efficient manner. Experimental evaluation with multiple real world datasets and user studies demonstrate that the efficacy of the proposed framework.
****

Learning Graph Models for Retrosynthesis Prediction
Vignesh Ram Somnath, Charlotte Bunne, Connor Coley, Andreas Krause, Regina Barzilay
Retrosynthesis prediction is a fundamental problem in organic synthesis, where the task is to identify precursor molecules that can be used to synthesize a target molecule. A key consideration in building neural models for this task is aligning model design with strategies adopted by chemists. Building on this viewpoint, this paper introduces a graph-based approach that capitalizes on the idea that the graph topology of precursor molecules is largely unaltered during a chemical reaction. The model first predicts the set of graph edits transforming the target into incomplete molecules called synthons. Next, the model learns to expand synthons into complete molecules by attaching relevant leaving groups. This decomposition simplifies the architecture, making its predictions more interpretable, and also amenable to manual correction. Our model achieves a top-1 accuracy of 53.7%, outperforming previous template-free and semi-template-based methods.
****

Differentiable Equilibrium Computation with Decision Diagrams for Stackelberg Models of Combinatorial Congestion Games

Shinsaku Sakaue, Kengo Nakamura
We address Stackelberg models of combinatorial congestion games (CCGs); we aim to optimize the parameters of CCGs so that the selfish behavior of non-atomic players attains desirable equilibria. This model is essential for designing such social infrastructures as traffic and communication networks. Nevertheless, computational approaches to the model have not been thoroughly studied due to two difficulties: (I) bilevel-programming structures and (II) the combinatorial nature of CCGs. We tackle them by carefully combining (I) the idea of \textit{differentiable} optimization and (II) data structures called \textit{zero-suppressed binary decision diagrams} (ZDDs), which can compactly represent sets of combinatorial strategies. Our algorithm numerically approximates the equilibria of CCGs, which we can differentiate with respect to parameters of CCGs by automatic differentiation. With the resulting derivatives, we can apply gradient-based methods to Stackelberg models of CCGs. Our method is tailored to induce Nesterov's acceleration and can fully utilize the empirical compactness of ZDDs. These technical advantages enable us to deal with CCGs with a vast number of combinatorial strategies. Experiments on real-world network design instances demonstrate the practicality of our method.
****
Inverse Optimal Control Adapted to the Noise Characteristics of the Human Sensorimotor System
Matthias Schultheis, Dominik Straub, Constantin A. Rothkopf
Computational level explanations based on optimal feedback control with signal-dependent noise have been able to account for a vast array of phenomena in human sensorimotor behavior. However, commonly a cost function needs to be assumed for a task and the optimality of human behavior is evaluated by comparing observed and predicted trajectories. Here, we introduce inverse optimal control with signal-dependent noise, which allows inferring the cost function from observed behavior. To do so, we formalize the problem as a partially observable Markov decision process and distinguish between the agent’s and the experimenter’s inference problems. Specifically, we derive a probabilistic formulation of the evolution of states and belief states and an approximation to the propagation equation in the linear-quadratic Gaussian problem with signal-dependent noise. We extend the model to the case of partial observability of state variables from the point of view of the experimenter. We show the feasibility of the approach through validation on synthetic data and application to experimental data. Our approach enables recovering the costs and benefits implicit in human sequential sensorimotor behavior, thereby reconciling normative and descriptive approaches in a computational framework.
****
Deep Neural Networks as Point Estimates for Deep Gaussian Processes
Vincent Dutordoir, James Hensman, Mark van der Wilk, Carl Henrik Ek, Zoubin Ghahramani, Nicolas Durrande
Neural networks and Gaussian processes are complementary in their strengths and weaknesses. Having a better understanding of their relationship comes with the promise to make each method benefit from the strengths of the other. In this work, we establish an equivalence between the forward passes of neural networks and (deep) sparse Gaussian process models. The theory we develop is based on interpreting activation functions as interdomain inducing features through a rigorous analysis of the interplay between activation functions and kernels. This results in models that can either be seen as neural networks with improved uncertainty prediction or deep Gaussian processes with increased prediction accuracy. These claims are supported by experimental results on regression and classification datasets.
****
Locality defeats the curse of dimensionality in convolutional teacher-student scenarios
Alessandro Favero, Francesco Cagnetta, Matthieu Wyart
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Causal Identification with Matrix Equations
Sanghack Lee, Elias Bareinboim
Causal effect identification is concerned with determining whether a causal effect is computable from a combination of qualitative assumptions about the underlying system (e.g., a causal graph) and distributions collected from this system. Many identification algorithms exclusively rely on graphical criteria made of a non-trivial combination of probability axioms, do-calculus, and refined c-factorization (e.g., Lee & Bareinboim, 2020). In a sequence of increasingly sophisticated results, it has been shown how proxy variables can be used to identify certain effects that would not be otherwise recoverable in challenging scenarios through solving matrix equations (e.g., Kuroki & Pearl, 2014; Miao et al., 2018). In this paper, we develop a new causal identification algorithm which utilizes both graphical criteria and matrix equations. Specifically, we first characterize the relationships between certain graphically-driven formulae and matrix multiplications. With such characterizations, we broaden the spectrum of proxy variable based identification conditions and further propose novel intermediary criteria based on the pseudoinverse of a matrix. Finally, we devise a causal effect identification algorithm, which accepts as input a collection of marginal, conditional, and interventional distributions, integrating enriched matrix-based criteria into a graphical identification approach.
****
Private and Non-private Uniformity Testing for Ranking Data
Róbert Busa-Fekete, Dimitris Fotakis, Emmanouil Zampetakis
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 Reinforcement Learning via Imagination with Derived Memory
Yao Mu, Yuzheng Zhuang, Bin Wang, Guangxiang Zhu, Wulong Liu, Jianyu Chen, Ping Luo, Shengbo Li, Chongjie Zhang, Jianye Hao
Model-based reinforcement learning aims to improve the sample efficiency of policy learning by modeling the dynamics of the environment. Recently, the latent dynamics model is further developed to enable fast planning in a compact space. It summarizes the high-dimensional experiences of an agent, which mimics the memory function of humans. Learning policies via imagination with the latent model shows great potential for solving complex tasks. However, only considering memories from the true experiences in the process of imagination could limit its advantages. Inspired by the memory prosthesis proposed by neuroscientists, we present a novel model-based reinforcement learning framework called Imagining with Derived Memory (IDM). It enables the agent to learn policy from enriched diverse imagination with prediction-reliability weight, thus improving sample efficiency and policy robustness. Experiments on various high-dimensional visual control tasks in the DMControl benchmark demonstrate that IDM outperforms previous state-of-the-art methods in terms of policy robustness and further improves the sample efficiency of the model-based method.
****
Compositional Transformers for Scene Generation
Dor Arad Hudson, Larry Zitnick
We introduce the GANformer2 model, an iterative object-oriented transformer, explored for the task of generative modeling. The network incorporates strong and explicit structural priors, to reflect the compositional nature of visual scenes, and synthesizes images through a sequential process. It operates in two stages: a fast and lightweight planning phase, where we draft a high-level scene layout, followed by an attention-based execution phase, where the layout is being refined, evolving into a rich and detailed picture. Our model moves away from conventional black-box GAN architectures that feature a flat and monolithic latent spa

ce towards a transparent design that encourages efficiency, controllability and interpretability. We demonstrate GANformer2's strengths and qualities through a careful evaluation over a range of datasets, from multi-object CLEVR scenes to the challenging COCO images, showing it successfully achieves state-of-the-art performance in terms of visual quality, diversity and consistency. Further experiments demonstrate the model's disentanglement and provide a deeper insight into its generative process, as it proceeds step-by-step from a rough initial sketch, to a detailed layout that accounts for objects' depths and dependencies, and up to the final high-resolution depiction of vibrant and intricate real-world scenes. See https://github.com/dorarad/gansformer for model implementation.
****
An Exponential Lower Bound for Linearly Realizable MDP with Constant Suboptimality Gap
Yuanhao Wang, Ruosong Wang, 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-authors prior to requesting a name change in the electronic proceedings.
****
Combating Noise: Semi-supervised Learning by Region Uncertainty Quantification
Zhenyu Wang, Ya-Li Li, Ye Guo, Shengjin Wang
Semi-supervised learning aims to leverage a large amount of unlabeled data for performance boosting. Existing works primarily focus on image classification. In this paper, we delve into semi-supervised learning for object detection, where labeled data are more labor-intensive to collect. Current methods are easily distracted by noisy regions generated by pseudo labels. To combat the noisy labeling, we propose noise-resistant semi-supervised learning by quantifying the region uncertainty. We first investigate the adverse effects brought by different forms of noise associated with pseudo labels. Then we propose to quantify the uncertainty of regions by identifying the noise-resistant properties of regions over different strengths. By importing the region uncertainty quantification and promoting multi-peak probability distribution output, we introduce uncertainty into training and further achieve noise-resistant learning. Experiments on both PASCAL VOC and MS COCO demonstrate the extraordinary performance of our method.
****
Reducing the Covariate Shift by Mirror Samples in Cross Domain Alignment
Yin Zhao, minquan wang, Longjun Cai
Eliminating the covariate shift cross domains is one of the common methods to deal with the issue of domain shift in visual unsupervised domain adaptation. However, current alignment methods, especially the prototype based or sample-level based methods neglect the structural properties of the underlying distribution and even break the condition of covariate shift. To relieve the limitations and conflicts, we introduce a novel concept named (virtual) mirror, which represents the equivalent sample in another domain. The equivalent sample pairs, named mirror pairs reflect the natural correspondence of the empirical distributions. Then a mirror loss, which aligns the mirror pairs cross domains, is constructed to enhance the alignment of the domains. The proposed method does not distort the internal structure of the underlying distribution. We also provide theoretical proof that the mirror samples and mirror loss have better asymptotic properties in reducing the domain shift. By applying the virtual mirror and mirror loss to the generic unsupervised domain adaptation model, we achieved consistently superior performance on several mainstream benchmarks.
****
Permutation-Invariant Variational Autoencoder for Graph-Level Representation Learning
Robin Winter, Frank Noe, Djork-Arné Clevert
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****
Causal Abstractions of Neural Networks
Atticus Geiger, Hanson Lu, Thomas Icard, Christopher Potts
Structural analysis methods (e.g., probing and feature attribution) are increasingly important tools for neural network analysis. We propose a new structural analysis method grounded in a formal theory of causal abstraction that provides rich characterizations of model-internal representations and their roles in input/output behavior. In this method, neural representations are aligned with variables in interpretable causal models, and then interchange interventions are used to experimentally verify that the neural representations have the causal properties of their aligned variables. We apply this method in a case study to analyze neural models trained on Multiply Quantified Natural Language Inference (MQNLI) corpus, a highly complex NLI dataset that was constructed with a tree-structured natural logic causal model. We discover that a BERT-based model with state-of-the-art performance successfully realizes parts of the natural logic model’s causal structure, whereas a simpler baseline model fails to show any such structure, demonstrating that neural representations encode the compositional structure of MQNLI examples.
****
Conic Blackwell Algorithm: Parameter-Free Convex-Concave Saddle-Point Solving
Julien Grand-Clément, Christian Kroer
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
3DP3: 3D Scene Perception via Probabilistic Programming
Nishad Gothoskar, Marco Cusumano-Towner, Ben Zinberg, Matin Ghavamizadeh, Falk Pollok, Austin Garrett, Josh Tenenbaum, Dan Gutfreund, Vikash Mansinghka
We present 3DP3, a framework for inverse graphics that uses inference in a structured generative model of objects, scenes, and images. 3DP3 uses (i) voxel models to represent the 3D shape of objects, (ii) hierarchical scene graphs to decompose scenes into objects and the contacts between them, and (iii) depth image likelihoods based on real-time graphics. Given an observed RGB-D image, 3DP3's inference algorithm infers the underlying latent 3D scene, including the object poses and a parsimonious joint parametrization of these poses, using fast bottom-up pose proposals, novel involutive MCMC updates of the scene graph structure, and, optionally, neural object detectors and pose estimators. We show that 3DP3 enables scene understanding that is aware of 3D shape, occlusion, and contact structure. Our results demonstrate that 3DP3 is more accurate at 6DoF object pose estimation from real images than deep learning baselines and shows better generalization to challenging scenes with novel viewpoints, contact, and partial observability.
****
Novel Upper Bounds for the Constrained Most Probable Explanation Task
Tahrima Rahman, Sara Rouhani, 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-authors prior to requesting a name change in the electronic proceedings.
****
Why Spectral Normalization Stabilizes GANs: Analysis and Improvements
Zinan Lin, Vyas Sekar, Giulia Fanti
Spectral normalization (SN) is a widely-used technique for improving the stability and sample quality of Generative Adversarial Networks (GANs). However, current understanding of SN's efficacy is limited. In this work, we show that SN controls two important failure modes of GAN training: exploding and vanishing gradients. Our proofs illustrate a (perhaps unintentional) connection with the successful LeCun initialization. This connection helps to explain why the most popular implementation of SN for GANs requires no hyper-parameter tuning, whereas stricte

r implementations of SN have poor empirical performance out-of-the-box. Unlike LeCun initialization which only controls gradient vanishing at the beginning of training, SN preserves this property throughout training. Building on this theoretical understanding, we propose a new spectral normalization technique: Bidirectional Scaled Spectral Normalization (BSSN), which incorporates insights from later improvements to LeCun initialization: Xavier initialization and Kaiming initialization. Theoretically, we show that BSSN gives better gradient control than SN. Empirically, we demonstrate that it outperforms SN in sample quality and training stability on several benchmark datasets.
****
$(\textrm{Implicit})^2$: Implicit Layers for Implicit Representations
Zhichun Huang, Shaojie Bai, J. Zico Kolter
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Mean-based Best Arm Identification in Stochastic Bandits under Reward Contamination
Arpan Mukherjee, Ali Tajer, Pin-Yu Chen, Payel Das
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
MADE: Exploration via Maximizing Deviation from Explored Regions
Tianjun Zhang, Paria Rashidinejad, Jiantao Jiao, Yuandong Tian, Joseph E. Gonzalez, Stuart Russell
In online reinforcement learning (RL), efficient exploration remains particularly challenging in high-dimensional environments with sparse rewards. In low-dimensional environments, where tabular parameterization is possible, count-based upper confidence bound (UCB) exploration methods achieve minimax near-optimal rates. However, it remains unclear how to efficiently implement UCB in realistic RL tasks that involve non-linear function approximation. To address this, we propose a new exploration approach via maximizing the deviation of the occupancy of the next policy from the explored regions. We add this term as an adaptive regularizer to the standard RL objective to balance exploration vs. exploitation. We pair the new objective with a provably convergent algorithm, giving rise to a new intrinsic reward that adjusts existing bonuses. The proposed intrinsic reward is easy to implement and combine with other existing RL algorithms to conduct exploration. As a proof of concept, we evaluate the new intrinsic reward on tabular examples across a variety of model-based and model-free algorithms, showing improvements over count-only exploration strategies. When tested on navigation and locomotion tasks from MiniGrid and DeepMind Control Suite benchmarks, our approach significantly improves sample efficiency over state-of-the-art methods.
****
Variational Automatic Curriculum Learning for Sparse-Reward Cooperative Multi-Agent Problems
Jiayu Chen, Yuanxin Zhang, Yuanfan Xu, Huimin Ma, Huazhong Yang, Jiaming Song, Yu Wang, Yi Wu
We introduce an automatic curriculum algorithm, Variational Automatic Curriculum Learning (VACL), for solving challenging goal-conditioned cooperative multi-agent reinforcement learning problems. We motivate our curriculum learning paradigm through a variational perspective, where the learning objective can be decomposed into two terms: task learning on the current curriculum, and curriculum update to a new task distribution. Local optimization over the second term suggests that the curriculum should gradually expand the training tasks from easy to hard. Our VACL algorithm implements this variational paradigm with two practical components, task expansion and entity curriculum, which produces a series of training tasks over both the task configurations as well as the number of entities in t

he task. Experiment results show that VACL solves a collection of sparse-reward problems with a large number of agents. Particularly, using a single desktop machine, VACL achieves 98% coverage rate with 100 agents in the simple-spread benchmark and reproduces the ramp-use behavior originally shown in OpenAI's hide-and-seek project.
****
Align before Fuse: Vision and Language Representation Learning with Momentum Distillation
Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, Steven Chu Hong Hoi
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Variational Model Inversion Attacks
Kuan-Chieh Wang, YAN FU, Ke Li, Ashish Khisti, Richard Zemel, Alireza Makhzani
Given the ubiquity of deep neural networks, it is important that these models do not reveal information about sensitive data that they have been trained on. In model inversion attacks, a malicious user attempts to recover the private dataset used to train a supervised neural network. A successful model inversion attack should generate realistic and diverse samples that accurately describe each of the classes in the private dataset. In this work, we provide a probabilistic interpretation of model inversion attacks, and formulate a variational objective that accounts for both diversity and accuracy. In order to optimize this variational objective, we choose a variational family defined in the code space of a deep generative model, trained on a public auxiliary dataset that shares some structural similarity with the target dataset. Empirically, our method substantially improves performance in terms of target attack accuracy, sample realism, and diversity on datasets of faces and chest X-ray images.
****
Graph Neural Networks with Adaptive Residual
Xiaorui Liu, Jiayuan Ding, Wei Jin, Han Xu, Yao Ma, Zitao Liu, Jiliang Tang
Graph neural networks (GNNs) have shown the power in graph representation learning for numerous tasks. In this work, we discover an interesting phenomenon that although residual connections in the message passing of GNNs help improve the performance, they immensely amplify GNNs' vulnerability against abnormal node features. This is undesirable because in real-world applications, node features in graphs could often be abnormal such as being naturally noisy or adversarially manipulated. We analyze possible reasons to understand this phenomenon and aim to design GNNs with stronger resilience to abnormal features. Our understandings motivate us to propose and derive a simple, efficient, interpretable, and adaptive message passing scheme, leading to a novel GNN with Adaptive Residual, AirGNN. Extensive experiments under various abnormal feature scenarios demonstrate the effectiveness of the proposed algorithm.
****
Efficient Active Learning for Gaussian Process Classification by Error Reduction
Guang Zhao, Edward Dougherty, Byung-Jun Yoon, Francis Alexander, Xiaoning Qian
Active learning sequentially selects the best instance for labeling by optimizing an acquisition function to enhance data/label efficiency. The selection can be either from a discrete instance set (pool-based scenario) or a continuous instance space (query synthesis scenario). In this work, we study both active learning scenarios for Gaussian Process Classification (GPC). The existing active learning strategies that maximize the Estimated Error Reduction (EER) aim at reducing the classification error after training with the new acquired instance in a one-step-look-ahead manner. The computation of EER-based acquisition functions is typically prohibitive as it requires retraining the GPC with every new query. Moreover, as the EER is not smooth, it can not be combined with gradient-based optimization techniques to efficiently explore the continuous instance space for query synthesis. To overcome these critical limitations, we develop computationall

y efficient algorithms for EER-based active learning with GPC. We derive the joint predictive distribution of label pairs as a one-dimensional integral, as a result of which the computation of the acquisition function avoids retraining the GPC for each query, remarkably reducing the computational overhead. We also derive the gradient chain rule to efficiently calculate the gradient of the acquisition function, which leads to the first query synthesis active learning algorithm implementing EER-based strategies. Our experiments clearly demonstrate the computational efficiency of the proposed algorithms. We also benchmark our algorithms on both synthetic and real-world datasets, which show superior performance in terms of sampling efficiency compared to the existing state-of-the-art algorithms.
****
Non-Asymptotic Analysis for Two Time-scale TDC with General Smooth Function Approximation
Yue Wang, Shaofeng Zou, Yi 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.
****
A Little Robustness Goes a Long Way: Leveraging Robust Features for Targeted Transfer Attacks
Jacob Springer, Melanie Mitchell, Garrett Kenyon
Adversarial examples for neural network image classifiers are known to be transferable: examples optimized to be misclassified by a source classifier are often misclassified as well by classifiers with different architectures. However, targeted adversarial examples—optimized to be classified as a chosen target class—tend to be less transferable between architectures. While prior research on constructing transferable targeted attacks has focused on improving the optimization procedure, in this work we examine the role of the source classifier. Here, we show that training the source classifier to be "slightly robust"—that is, robust to small-magnitude adversarial examples—substantially improves the transferability of class-targeted and representation-targeted adversarial attacks, even between architectures as different as convolutional neural networks and transformers. The results we present provide insight into the nature of adversarial examples as well as the mechanisms underlying so-called "robust" classifiers.
****
TriBERT: Human-centric Audio-visual Representation Learning
Tanzila Rahman, Mengyu Yang, Leonid Sigal
The recent success of transformer models in language, such as BERT, has motivated the use of such architectures for multi-modal feature learning and tasks. However, most multi-modal variants (e.g., ViLBERT) have limited themselves to visual-linguistic data. Relatively few have explored its use in audio-visual modalities, and none, to our knowledge, illustrate them in the context of granular audio-visual detection or segmentation tasks such as sound source separation and localization. In this work, we introduce TriBERT -- a transformer-based architecture, inspired by ViLBERT, which enables contextual feature learning across three modalities: vision, pose, and audio, with the use of flexible co-attention. The use of pose keypoints is inspired by recent works that illustrate that such representations can significantly boost performance in many audio-visual scenarios where often one or more persons are responsible for the sound explicitly (e.g., talking) or implicitly (e.g., sound produced as a function of human manipulating an object). From a technical perspective, as part of the TriBERT architecture, we introduce a learned visual tokenization scheme based on spatial attention and leverage weak-supervision to allow granular cross-modal interactions for visual and pose modalities. Further, we supplement learning with sound-source separation loss formulated across all three streams. We pre-train our model on the large MUSIC21 dataset and demonstrate improved performance in audio-visual sound source separation on that dataset as well as other datasets through fine-tuning. In addition, we show that the learned TriBERT representations are generic and significa

ntly improve performance on other audio-visual tasks such as cross-modal audio-visual-pose retrieval by as much as 66.7% in top-1 accuracy.
****

How does a Neural Network's Architecture Impact its Robustness to Noisy Labels?
Jingling Li, Mozhi Zhang, Keyulu Xu, John Dickerson, Jimmy Ba
Noisy labels are inevitable in large real-world datasets. In this work, we explore an area understudied by previous works --- how the network's architecture impacts its robustness to noisy labels. We provide a formal framework connecting the robustness of a network to the alignments between its architecture and target/noise functions. Our framework measures a network's robustness via the predictive power in its representations --- the test performance of a linear model trained on the learned representations using a small set of clean labels. We hypothesize that a network is more robust to noisy labels if its architecture is more aligned with the target function than the noise. To support our hypothesis, we provide both theoretical and empirical evidence across various neural network architectures and different domains. We also find that when the network is well-aligned with the target function, its predictive power in representations could improve upon state-of-the-art (SOTA) noisy-label-training methods in terms of test accuracy and even outperform sophisticated methods that use clean labels.
****

Calibration and Consistency of Adversarial Surrogate Losses
Pranjal Awasthi, Natalie Frank, Anqi Mao, Mehryar Mohri, Yutao Zhong
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

The Value of Information When Deciding What to Learn
Dilip Arumugam, Benjamin Van Roy
All sequential decision-making agents explore so as to acquire knowledge about a particular target. It is often the responsibility of the agent designer to construct this target which, in rich and complex environments, constitutes a onerous burden; without full knowledge of the environment itself, a designer may forge a sub-optimal learning target that poorly balances the amount of information an agent must acquire to identify the target against the target's associated performance shortfall. While recent work has developed a connection between learning targets and rate-distortion theory to address this challenge and empower agents that decide what to learn in an automated fashion, the proposed algorithm does not optimally tackle the equally important challenge of efficient information acquisition. In this work, building upon the seminal design principle of information-directed sampling (Russo & Van Roy, 2014), we address this shortcoming directly to couple optimal information acquisition with the optimal design of learning targets. Along the way, we offer new insights into learning targets from the literature on rate-distortion theory before turning to empirical results that confirm the value of information when deciding what to learn.
****

Co-Adaptation of Algorithmic and Implementational Innovations in Inference-based Deep Reinforcement Learning
Hiroki Furuta, Tadashi Kozuno, Tatsuya Matsushima, Yutaka Matsuo, Shixiang (Shane) Gu
Recently many algorithms were devised for reinforcement learning (RL) with function approximation. While they have clear algorithmic distinctions, they also have many implementation differences that are algorithm-independent and sometimes under-emphasized. Such mixing of algorithmic novelty and implementation craftsmanship makes rigorous analyses of the sources of performance improvements across algorithms difficult. In this work, we focus on a series of off-policy inference-based actor-critic algorithms -- MPO, AWR, and SAC -- to decouple their algorithmic innovations and implementation decisions. We present unified derivations through a single control-as-inference objective, where we can categorize each algorithm as based on either Expectation-Maximization (EM) or direct Kullback-Leibler

 (KL) divergence minimization and treat the rest of specifications as implementa tion details. We performed extensive ablation studies, and identified substantia l performance drops whenever implementation details are mismatched for algorithm ic choices. These results show which implementation or code details are co-adapt ed and co-evolved with algorithms, and which are transferable across algorithms: as examples, we identified that tanh Gaussian policy and network sizes are high ly adapted to algorithmic types, while layer normalization and ELU are critical for MPO's performances but also transfer to noticeable gains in SAC. We hope our work can inspire future work to further demystify sources of performance improv ements across multiple algorithms and allow researchers to build on one another' s both algorithmic and implementational innovations.
****
Can fMRI reveal the representation of syntactic structure in the brain?
Aniketh Janardhan Reddy, Leila Wehbe
While studying semantics in the brain, neuroscientists use two approaches. One i s to identify areas that are correlated with semantic processing load. Another i s to find areas that are predicted by the semantic representation of the stimulu s words. However, most studies of syntax have focused only on identifying areas correlated with syntactic processing load. One possible reason for this discrep ancy is that representing syntactic structure in an embedding space such that it can be used to model brain activity is a non-trivial computational problem. Ano ther possible reason is that it is unclear if the low signal-to-noise ratio of n euroimaging tools such as functional Magnetic Resonance Imaging (fMRI) can allow us to reveal the correlates of complex (and perhaps subtle) syntactic represent ations. In this study, we propose novel multi-dimensional features that encode i nformation about the syntactic structure of sentences. Using these features and fMRI recordings of participants reading a natural text, we model the brain repre sentation of syntax. First, we find that our syntactic structure-based features explain additional variance in the brain activity of various parts of the langua ge system, even after controlling for complexity metrics that capture processing load. At the same time, we see that regions well-predicted by syntactic feature s are distributed in the language system and are not distinguishable from those processing semantics. Our code and data will be available at https://github.com/ anikethjr/brainsyntacticrepresentations.
****
Robust Implicit Networks via Non-Euclidean Contractions
Saber Jafarpour, Alexander Davydov, Anton Proskurnikov, Francesco Bullo
Requests for name changes in the electronic proceedings will be accepted with 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 Kernel-based Test of Independence for Cluster-correlated Data
Hongjiao Liu, Anna Plantinga, Yunhua Xiang, Michael Wu
The Hilbert-Schmidt Independence Criterion (HSIC) is a powerful kernel-based sta tistic for assessing the generalized dependence between two multivariate variabl es. However, independence testing based on the HSIC is not directly possible for cluster-correlated data. Such a correlation pattern among the observations aris es in many practical situations, e.g., family-based and longitudinal data, and r equires proper accommodation. Therefore, we propose a novel HSIC-based independe nce test to evaluate the dependence between two multivariate variables based on cluster-correlated data. Using the previously proposed empirical HSIC as our tes t statistic, we derive its asymptotic distribution under the null hypothesis of independence between the two variables but in the presence of sample correlation . Based on both simulation studies and real data analysis, we show that, with cl ustered data, our approach effectively controls type I error and has a higher st atistical power than competing methods.
****
Efficient methods for Gaussian Markov random fields under sparse linear constrai nts

David Bolin, Jonas Wallin
Methods for inference and simulation of linearly constrained Gaussian Markov Ran dom Fields (GMRF) are computationally prohibitive when the number of constraints is large. In some cases, such as for intrinsic GMRFs, they may even be unfeasib le. We propose a new class of methods to overcome these challenges in the common case of sparse constraints, where one has a large number of constraints and eac h only involves a few elements. Our methods rely on a basis transformation into blocks of constrained versus non-constrained subspaces, and we show that the met hods greatly outperform existing alternatives in terms of computational cost. By combining the proposed methods with the stochastic partial differential equatio n approach for Gaussian random fields, we also show how to formulate Gaussian pr ocess regression with linear constraints in a GMRF setting to reduce computation al cost. This is illustrated in two applications with simulated data.
****

Sparse is Enough in Scaling Transformers
Sebastian Jaszczur, Aakanksha Chowdhery, Afroz Mohiuddin, LUKASZ KAISER, Wojciec h Gajewski, Henryk Michalewski, Jonni Kanerva
Large Transformer models yield impressive results on many tasks, but are expensi ve to train, or even fine-tune, and so slow at decoding that their use and study becomes out of reach. We address this problem by leveraging sparsity. We study sparse variants for all layers in the Transformer and propose Scaling Transforme rs, a family of next generation Transformer models that use sparse layers to sca le efficiently and perform unbatched decoding much faster than the standard Tran sformer as we scale up the model size. Surprisingly, the sparse layers are enoug h to obtain the same perplexity as the standard Transformer with the same number of parameters. We also integrate with prior sparsity approaches to attention an d enable fast inference on long sequences even with limited memory. This results in performance competitive to the state-of-the-art on long text summarization.
****

Sparse Training via Boosting Pruning Plasticity with Neuroregeneration
Shiwei Liu, Tianlong Chen, Xiaohan Chen, Zahra Atashgahi, Lu Yin, Huanyu Kou, Li Shen, Mykola Pechenizkiy, Zhangyang Wang, Decebal Constantin Mocanu
Works on lottery ticket hypothesis (LTH) and single-shot network pruning (SNIP) have raised a lot of attention currently on post-training pruning (iterative mag nitude pruning), and before-training pruning (pruning at initialization). The fo rmer method suffers from an extremely large computation cost and the latter usua lly struggles with insufficient performance. In comparison, during-training prun ing, a class of pruning methods that simultaneously enjoys the training/inferenc e efficiency and the comparable performance, temporarily, has been less explored . To better understand during-training pruning, we quantitatively study the effe ct of pruning throughout training from the perspective of pruning plasticity (th e ability of the pruned networks to recover the original performance). Pruning p lasticity can help explain several other empirical observations about neural net work pruning in literature. We further find that pruning plasticity can be subst antially improved by injecting a brain-inspired mechanism called neuroregenerati on, i.e., to regenerate the same number of connections as pruned. We design a no vel gradual magnitude pruning (GMP) method, named gradual pruning with zero-cost neuroregeneration (GraNet), that advances state of the art. Perhaps most impres sively, its sparse-to-sparse version for the first time boosts the sparse-to-spa rse training performance over various dense-to-sparse methods with ResNet-50 on ImageNet without extending the training time. We release all codes in https://gi thub.com/Shiweiliuiiiiiii/GraNet.
****

Low-Fidelity Video Encoder Optimization for Temporal Action Localization
Mengmeng Xu, Juan Manuel Perez Rua, Xiatian Zhu, Bernard Ghanem, Brais Martinez
Most existing temporal action localization (TAL) methods rely on a transfer lear ning pipeline: by first optimizing a video encoder on a large action classificat ion dataset (i.e., source domain), followed by freezing the encoder and training a TAL head on the action localization dataset (i.e., target domain). This resul ts in a task discrepancy problem for the video encoder – trained for action clas

sification, but used for TAL. Intuitively, joint optimization with both the video encoder and TAL head is a strong baseline solution to this discrepancy. However, this is not operable for TAL subject to the GPU memory constraints, due to the prohibitive computational cost in processing long untrimmed videos. In this paper, we resolve this challenge by introducing a novel low-fidelity (LoFi) video encoder optimization method. Instead of always using the full training configurations in TAL learning, we propose to reduce the mini-batch composition in terms of temporal, spatial, or spatio-temporal resolution so that jointly optimizing the video encoder and TAL head becomes operable under the same memory conditions of a mid-range hardware budget. Crucially, this enables the gradients to flow backwards through the video encoder conditioned on a TAL supervision loss, favourably solving the task discrepancy problem and providing more effective feature representations. Extensive experiments show that the proposed LoFi optimization approach can significantly enhance the performance of existing TAL methods. Encouragingly, even with a lightweight ResNet18 based video encoder in a single RGB stream, our method surpasses two-stream (RGB + optical-flow) ResNet50 based alternatives, often by a good margin. Our code is publicly available at https://github.com/saic-fi/lofiactionlocalization.
****
On Provable Benefits of Depth in Training Graph Convolutional Networks
Weilin Cong, Morteza Ramezani, Mehrdad Mahdavi
Graph Convolutional Networks (GCNs) are known to suffer from performance degradation as the number of layers increases, which is usually attributed to over-smoothing. Despite the apparent consensus, we observe that there exists a discrepancy between the theoretical understanding of over-smoothing and the practical capabilities of GCNs. Specifically, we argue that over-smoothing does not necessarily happen in practice, a deeper model is provably expressive, can converge to global optimum with linear convergence rate, and achieve very high training accuracy as long as properly trained. Despite being capable of achieving high training accuracy, empirical results show that the deeper models generalize poorly on the testing stage and existing theoretical understanding of such behavior remains elusive. To achieve better understanding, we carefully analyze the generalization capability of GCNs, and show that the training strategies to achieve high training accuracy significantly deteriorate the generalization capability of GCNs. Motivated by these findings, we propose a decoupled structure for GCNs that detaches weight matrices from feature propagation to preserve the expressive power and ensure good generalization performance. We conduct empirical evaluations on various synthetic and real-world datasets to validate the correctness of our theory.
****
Practical Near Neighbor Search via Group Testing
Joshua Engels, Benjamin Coleman, Anshumali Shrivastava
We present a new algorithm for the approximate near neighbor problem that combines classical ideas from group testing with locality-sensitive hashing (LSH). We reduce the near neighbor search problem to a group testing problem by designating neighbors as "positives," non-neighbors as "negatives," and approximate membership queries as group tests. We instantiate this framework using distance-sensitive Bloom Filters to Identify Near-Neighbor Groups (FLINNG). We prove that FLINNG has sub-linear query time and show that our algorithm comes with a variety of practical advantages. For example, FLINNG can be constructed in a single pass through the data, consists entirely of efficient integer operations, and does not require any distance computations. We conduct large-scale experiments on high-dimensional search tasks such as genome search, URL similarity search, and embedding search over the massive YFCC100M dataset. In our comparison with leading algorithms such as HNSW and FAISS, we find that FLINNG can provide up to a 10x query speedup with substantially smaller indexing time and memory.
****
Baby Intuitions Benchmark (BIB): Discerning the goals, preferences, and actions of others
Kanishk Gandhi, Gala Stojnic, Brenden M. Lake, Moira R Dillon

To achieve human-like common sense about everyday life, machine learning systems must understand and reason about the goals, preferences, and actions of other agents in the environment. By the end of their first year of life, human infants intuitively achieve such common sense, and these cognitive achievements lay the foundation for humans' rich and complex understanding of the mental states of others. Can machines achieve generalizable, commonsense reasoning about other agents like human infants? The Baby Intuitions Benchmark (BIB) challenges machines to predict the plausibility of an agent's behavior based on the underlying causes of its actions. Because BIB's content and paradigm are adopted from developmental cognitive science, BIB allows for direct comparison between human and machine performance. Nevertheless, recently proposed, deep-learning-based agency reasoning models fail to show infant-like reasoning, leaving BIB an open challenge.
****

Neural Hybrid Automata: Learning Dynamics With Multiple Modes and Stochastic Transitions

Michael Poli, Stefano Massaroli, Luca Scimeca, Sanghyuk Chun, Seong Joon Oh, Atsushi Yamashita, Hajime Asama, Jinkyoo Park, Animesh Garg

Effective control and prediction of dynamical systems require appropriate handling of continuous-time and discrete, event-triggered processes. Stochastic hybrid systems (SHSs), common across engineering domains, provide a formalism for dynamical systems subject to discrete, possibly stochastic, state jumps and multi-modal continuous-time flows. Despite the versatility and importance of SHSs across applications, a general procedure for the explicit learning of both discrete events and multi-mode continuous dynamics remains an open problem. This work introduces Neural Hybrid Automata (NHAs), a recipe for learning SHS dynamics without a priori knowledge on the number, mode parameters, and inter-modal transition dynamics. NHAs provide a systematic inference method based on normalizing flows, neural differential equations, and self-supervision. We showcase NHAs on several tasks, including mode recovery and flow learning in systems with stochastic transitions, and end-to-end learning of hierarchical robot controllers.
****

Fast Projection onto the Capped Simplex with Applications to Sparse Regression in Bioinformatics

Man Shun Ang, Jianzhu Ma, Nianjun Liu, Kun Huang, Yijie Wang

We consider the problem of projecting a vector onto the so-called k-capped simplex, which is a hyper-cube cut by a hyperplane.For an n-dimensional input vector with bounded elements, we found that a simple algorithm based on Newton's method is able to solve the projection problem to high precision with a complexity roughly about O(n), which has a much lower computational cost compared with the existing sorting-based methods proposed in the literature.We provide a theory for partial explanation and justification of the method.We demonstrate that the proposed algorithm can produce a solution of the projection problem with high precision on large scale datasets, and the algorithm is able to significantly outperform the state-of-the-art methods in terms of runtime (about 6-8 times faster than a commercial software with respect to CPU time for input vector with 1 million variables or more).We further illustrate the effectiveness of the proposed algorithm on solving sparse regression in a bioinformatics problem.Empirical results on the GWAS dataset (with 1,500,000 single-nucleotide polymorphisms) show that, when using the proposed method to accelerate the Projected Quasi-Newton (PQN) method, the accelerated PQN algorithm is able to handle huge-scale regression problem and it is more efficient (about 3-6 times faster) than the current state-of-the-art methods.
****

The Many Faces of Adversarial Risk

Muni Sreenivas Pydi, Varun Jog

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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-Adaptive Nonlinear Control: Theory and Algorithms
Guanya Shi, Kamyar Azizzadenesheli, Michael O'Connell, Soon-Jo Chung, Yisong Yue
We present an online multi-task learning approach for adaptive nonlinear control, which we call Online Meta-Adaptive Control (OMAC). The goal is to control a nonlinear system subject to adversarial disturbance and unknown \emph{environment-dependent} nonlinear dynamics, under the assumption that the environment-dependent dynamics can be well captured with some shared representation. Our approach is motivated by robot control, where a robotic system encounters a sequence of new environmental conditions that it must quickly adapt to. A key emphasis is to integrate online representation learning with established methods from control theory, in order to arrive at a unified framework that yields both control-theoretic and learning-theoretic guarantees. We provide instantiations of our approach under varying conditions, leading to the first non-asymptotic end-to-end convergence guarantee for multi-task nonlinear control. OMAC can also be integrated with deep representation learning. Experiments show that OMAC significantly outperforms conventional adaptive control approaches which do not learn the shared representation, in inverted pendulum and 6-DoF drone control tasks under varying wind conditions.
****
Compositional Reinforcement Learning from Logical Specifications
Kishor Jothimurugan, Suguman Bansal, Osbert Bastani, Rajeev Alur
We study the problem of learning control policies for complex tasks given by logical specifications. Recent approaches automatically generate a reward function from a given specification and use a suitable reinforcement learning algorithm to learn a policy that maximizes the expected reward. These approaches, however, scale poorly to complex tasks that require high-level planning. In this work, we develop a compositional learning approach, called DIRL, that interleaves high-level planning and reinforcement learning. First, DIRL encodes the specification as an abstract graph; intuitively, vertices and edges of the graph correspond to regions of the state space and simpler sub-tasks, respectively. Our approach then incorporates reinforcement learning to learn neural network policies for each edge (sub-task) within a Dijkstra-style planning algorithm to compute a high-level plan in the graph. An evaluation of the proposed approach on a set of challenging control benchmarks with continuous state and action spaces demonstrates that it outperforms state-of-the-art baselines.
****
Differentiable Quality Diversity
Matthew Fontaine, Stefanos Nikolaidis
Quality diversity (QD) is a growing branch of stochastic optimization research that studies the problem of generating an archive of solutions that maximize a given objective function but are also diverse with respect to a set of specified measure functions. However, even when these functions are differentiable, QD algorithms treat them as "black boxes", ignoring gradient information. We present the differentiable quality diversity (DQD) problem, a special case of QD, where both the objective and measure functions are first order differentiable. We then present MAP-Elites via a Gradient Arborescence (MEGA), a DQD algorithm that leverages gradient information to efficiently explore the joint range of the objective and measure functions. Results in two QD benchmark domains and in searching the latent space of a StyleGAN show that MEGA significantly outperforms state-of-the-art QD algorithms, highlighting DQD's promise for efficient quality diversity optimization when gradient information is available. Source code is available at https://github.com/icaros-usc/dqd.
****
Credit Assignment Through Broadcasting a Global Error Vector
David Clark, L F Abbott, Sueyeon Chung
Backpropagation (BP) uses detailed, unit-specific feedback to train deep neural networks (DNNs) with remarkable success. That biological neural circuits appear to perform credit assignment, but cannot implement BP, implies the existence of other powerful learning algorithms. Here, we explore the extent to which a globally broadcast learning signal, coupled with local weight updates, enables traini

ng of DNNs. We present both a learning rule, called global error-vector broadcas
ting (GEVB), and a class of DNNs, called vectorized nonnegative networks (VNNs),
 in which this learning rule operates. VNNs have vector-valued units and nonnega
tive weights past the first layer. The GEVB learning rule generalizes three-fact
or Hebbian learning, updating each weight by an amount proportional to the inner
 product of the presynaptic activation and a globally broadcast error vector whe
n the postsynaptic unit is active. We prove that these weight updates are matche
d in sign to the gradient, enabling accurate credit assignment. Moreover, at ini
tialization, these updates are exactly proportional to the gradient in the limit
 of infinite network width. GEVB matches the performance of BP in VNNs, and in s
ome cases outperforms direct feedback alignment (DFA) applied in conventional ne
tworks. Unlike DFA, GEVB successfully trains convolutional layers. Altogether, o
ur theoretical and empirical results point to a surprisingly powerful role for a
 global learning signal in training DNNs.
****
An Online Method for A Class of Distributionally Robust Optimization with Non-co
nvex Objectives
Qi Qi, Zhishuai Guo, Yi Xu, Rong Jin, Tianbao Yang
In this paper, we propose a practical online method for solving a class of distr
ibutional robust optimization (DRO) with non-convex objectives, which has import
ant applications in machine learning for improving the robustness of neural netw
orks. In the literature, most methods for solving DRO are based on stochastic pr
imal-dual methods. However, primal-dual methods for DRO suffer from several draw
backs: (1) manipulating a high-dimensional dual variable corresponding to the si
ze of data is time expensive; (2) they are not friendly to online learning where
 data is coming sequentially. To address these issues, we consider a class of DR
O with an KL divergence regularization on the dual variables, transform the min-
max problem into a compositional minimization problem, and propose practical dua
lity-free online stochastic methods without requiring a large mini-batch size. W
e establish the state-of-the-art complexities of the proposed methods with and w
ithout a Polyak-■ojasiewicz (PL) condition of the objective. Empirical studies o
n large-scale deep learning tasks (i) demonstrate that our method can speed up t
he training by more than 2 times than baseline methods and save days of training
 time on a large-scale dataset with ~ 265K images, and (ii) verify the supreme p
erformance of DRO over Empirical Risk Minimization (ERM) on imbalanced datasets.
 Of independent interest, the proposed method can be also used for solving a fam
ily of stochastic compositional problems with state-of-the-art complexities.
****
A single gradient step finds adversarial examples on random two-layers neural ne
tworks
Sebastien Bubeck, Yeshwanth Cherapanamjeri, Gauthier Gidel, Remi Tachet des Comb
es
Daniely and Schacham recently showed that gradient descent finds adversarial exa
mples on random undercomplete two-layers ReLU neural networks. The term “underco
mplete” refers to the fact that their proof only holds when the number of neuron
s is a vanishing fraction of the ambient dimension. We extend their result to th
e overcomplete case, where the number of neurons is larger than the dimension (y
et also subexponential in the dimension). In fact we prove that a single step of
 gradient descent suffices. We also show this result for any subexponential widt
h random neural network with smooth activation function.
****
Parameterized Knowledge Transfer for Personalized Federated Learning
Jie Zhang, Song Guo, Xiaosong Ma, Haozhao Wang, Wenchao Xu, Feijie Wu
In recent years, personalized federated learning (pFL) has attracted increasing
attention for its potential in dealing with statistical heterogeneity among clie
nts. However, the state-of-the-art pFL methods rely on model parameters aggregat
ion at the server side, which require all models to have the same structure and
size, and thus limits the application for more heterogeneous scenarios. To deal
with such model constraints, we exploit the potentials of heterogeneous model se
ttings and propose a novel training framework to employ personalized models for

different clients. Specifically, we formulate the aggregation procedure in original pFL into a personalized group knowledge transfer training algorithm, namely, KT-pFL, which enables each client to maintain a personalized soft prediction at the server side to guide the others' local training. KT-pFL updates the personalized soft prediction of each client by a linear combination of all local soft predictions using a knowledge coefficient matrix, which can adaptively reinforce the collaboration among clients who own similar data distribution. Furthermore, to quantify the contributions of each client to others' personalized training, the knowledge coefficient matrix is parameterized so that it can be trained simultaneously with the models. The knowledge coefficient matrix and the model parameters are alternatively updated in each round following the gradient descent way. Extensive experiments on various datasets (EMNIST, Fashion_MNIST, CIFAR-10) are conducted under different settings (heterogeneous models and data distributions). It is demonstrated that the proposed framework is the first federated learning paradigm that realizes personalized model training via parameterized group knowledge transfer while achieving significant performance gain comparing with state-of-the-art algorithms.
****
Contrastively Disentangled Sequential Variational Autoencoder
Junwen Bai, Weiran Wang, Carla P. Gomes
Self-supervised disentangled representation learning is a critical task in sequence modeling. The learnt representations contribute to better model interpretability as well as the data generation, and improve the sample efficiency for downstream tasks. We propose a novel sequence representation learning method, named Contrastively Disentangled Sequential Variational Autoencoder (C-DSVAE), to extract and separate the static (time-invariant) and dynamic (time-variant) factors in the latent space. Different from previous sequential variational autoencoder methods, we use a novel evidence lower bound which maximizes the mutual information between the input and the latent factors, while penalizes the mutual information between the static and dynamic factors. We leverage contrastive estimations of the mutual information terms in training, together with simple yet effective augmentation techniques, to introduce additional inductive biases. Our experiments show that C-DSVAE significantly outperforms the previous state-of-the-art methods on multiple metrics.
****
Recursive Causal Structure Learning in the Presence of Latent Variables and Selection Bias
Sina Akbari, Ehsan Mokhtarian, AmirEmad Ghassami, Negar Kiyavash
We consider the problem of learning the causal MAG of a system from observational data in the presence of latent variables and selection bias. Constraint-based methods are one of the main approaches for solving this problem, but the existing methods are either computationally impractical when dealing with large graphs or lacking completeness guarantees. We propose a novel computationally efficient recursive constraint-based method that is sound and complete. The key idea of our approach is that at each iteration a specific type of variable is identified and removed. This allows us to learn the structure efficiently and recursively, as this technique reduces both the number of required conditional independence (CI) tests and the size of the conditioning sets. The former substantially reduces the computational complexity, while the latter results in more reliable CI tests. We provide an upper bound on the number of required CI tests in the worst case. To the best of our knowledge, this is the tightest bound in the literature. We further provide a lower bound on the number of CI tests required by any constraint-based method. The upper bound of our proposed approach and the lower bound at most differ by a factor equal to the number of variables in the worst case. We provide experimental results to compare the proposed approach with the state of the art on both synthetic and real-world structures.
****
Generalization Error Rates in Kernel Regression: The Crossover from the Noiseless to Noisy Regime
Hugo Cui, Bruno Loureiro, Florent Krzakala, Lenka Zdeborová

In this manuscript we consider Kernel Ridge Regression (KRR) under the Gaussian
design. Exponents for the decay of the excess generalization error of KRR have b
een reported in various works under the assumption of power-law decay of eigenva
lues of the features co-variance. These decays were, however, provided for sizea
bly different setups, namely in the noiseless case with constant regularization
and in the noisy optimally regularized case. Intermediary settings have been lef
t substantially uncharted. In this work, we unify and extend this line of work,
providing characterization of all regimes and excess error decay rates that can
be observed in terms of the interplay of noise and regularization. In particular
, we show the existence of a transition in the noisy setting between the noisele
ss exponents to its noisy values as the sample complexity is increased. Finally,
 we illustrate how this crossover can also be observed on real data sets.
****
Learning Gaussian Mixtures with Generalized Linear Models: Precise Asymptotics i
n High-dimensions
Bruno Loureiro, Gabriele Sicuro, Cedric Gerbelot, Alessandro Pacco, Florent Krza
kala, 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.
****
Spectral embedding for dynamic networks with stability guarantees
Ian Gallagher, Andrew Jones, Patrick Rubin-Delanchy
We consider the problem of embedding a dynamic network, to obtain time-evolving
vector representations of each node, which can then be used to describe changes
in behaviour of individual nodes, communities, or the entire graph. Given this o
pen-ended remit, we argue that two types of stability in the spatio-temporal pos
itioning of nodes are desirable: to assign the same position, up to noise, to no
des behaving similarly at a given time (cross-sectional stability) and a constan
t position, up to noise, to a single node behaving similarly across different ti
mes (longitudinal stability). Similarity in behaviour is defined formally using
notions of exchangeability under a dynamic latent position network model. By sho
wing how this model can be recast as a multilayer random dot product graph, we d
emonstrate that unfolded adjacency spectral embedding satisfies both stability c
onditions. We also show how two alternative methods, omnibus and independent spe
ctral embedding, alternately lack one or the other form of stability.
****
Infinite Time Horizon Safety of Bayesian Neural Networks
Mathias Lechner, ■or■e Žikeli■, Krishnendu Chatterjee, Thomas Henzinger
Bayesian neural networks (BNNs) place distributions over the weights of a neural
 network to model uncertainty in the data and the network's prediction.We consid
er the problem of verifying safety when running a Bayesian neural network policy
 in a feedback loop with infinite time horizon systems.Compared to the existing
sampling-based approaches, which are inapplicable to the infinite time horizon s
etting, we train a separate deterministic neural network that serves as an infin
ite time horizon safety certificate.In particular, we show that the certificate
network guarantees the safety of the system over a subset of the BNN weight post
erior's support. Our method first computes a safe weight set and then alters the
 BNN's weight posterior to reject samples outside this set. Moreover, we show ho
w to extend our approach to a safe-exploration reinforcement learning setting, i
n order to avoid unsafe trajectories during the training of the policy. We evalu
ate our approach on a series of reinforcement learning benchmarks, including non
-Lyapunovian safety specifications.
****
Towards understanding retrosynthesis by energy-based models
Ruoxi Sun, Hanjun Dai, Li Li, Steven Kearnes, Bo Dai
Retrosynthesis is the process of identifying a set of reactants to synthesize a
target molecule. It is of vital importance to material design and drug discovery
. Existing machine learning approaches based on language models and graph neural

networks have achieved encouraging results. However, the inner connections of these models are rarely discussed, and rigorous evaluations of these models are largely in need. In this paper, we propose a framework that unifies sequence- and graph-based methods as energy-based models (EBMs) with different energy functions. This unified view establishes connections and reveals the differences between models, thereby enhancing our understanding of model design. We also provide a comprehensive assessment of performance to the community. Moreover, we present a novel dual variant within the framework that performs consistent training to induce the agreement between forward- and backward-prediction. This model improves the state-of-the-art of template-free methods with or without reaction types.

****

List-Decodable Mean Estimation in Nearly-PCA Time

Ilias Diakonikolas, Daniel Kane, Daniel Kongsgaard, 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.

****

Distributed Zero-Order Optimization under Adversarial Noise

Arya Akhavan, Massimiliano Pontil, Alexandre Tsybakov

We study the problem of distributed zero-order optimization for a class of strongly convex functions. They are formed by the average of local objectives, associated to different nodes in a prescribed network. We propose a distributed zero-order projected gradient descent algorithm to solve the problem. Exchange of information within the network is permitted only between neighbouring nodes. An important feature of our procedure is that it can query only function values, subject to a general noise model, that does not require zero mean or independent errors. We derive upper bounds for the average cumulative regret and optimization error of the algorithm which highlight the role played by a network connectivity parameter, the number of variables, the noise level, the strong convexity parameter, and smoothness properties of the local objectives. The bounds indicate some key improvements of our method over the state-of-the-art, both in the distributed and standard zero-order optimization settings.

****

Reliable Estimation of KL Divergence using a Discriminator in Reproducing Kernel Hilbert Space

Sandesh Ghimire, Aria Masoomi, Jennifer Dy

Estimating Kullback–Leibler (KL) divergence from samples of two distributions is essential in many machine learning problems. Variational methods using neural network discriminator have been proposed to achieve this task in a scalable manner. However, we noticed that most of these methods using neural network discriminators suffer from high fluctuations (variance) in estimates and instability in training. In this paper, we look at this issue from statistical learning theory and function space complexity perspective to understand why this happens and how to solve it. We argue that the cause of these pathologies is lack of control over the complexity of the neural network discriminator function and could be mitigated by controlling it. To achieve this objective, we 1) present a novel construction of the discriminator in the Reproducing Kernel Hilbert Space (RKHS), 2) theoretically relate the error probability bound of the KL estimates to the complexity of the discriminator in the RKHS space, 3) present a scalable way to control the complexity (RKHS norm) of the discriminator for a reliable estimation of KL divergence, and 4) prove the consistency of the proposed estimator. In three different applications of KL divergence -- estimation of KL, estimation of mutual information and Variational Bayes -- we show that by controlling the complexity as developed in the theory, we are able to reduce the variance of KL estimates and stabilize the training.

****

Latent Matters: Learning Deep State-Space Models

Alexej Klushyn, Richard Kurle, Maximilian Soelch, Botond Cseke, Patrick van der

Smagt
Deep state-space models (DSSMs) enable temporal predictions by learning the underlying dynamics of observed sequence data. They are often trained by maximising the evidence lower bound. However, as we show, this does not ensure the model actually learns the underlying dynamics. We therefore propose a constrained optimisation framework as a general approach for training DSSMs. Building upon this, we introduce the extended Kalman VAE (EKVAE), which combines amortised variational inference with classic Bayesian filtering/smoothing to model dynamics more accurately than RNN-based DSSMs. Our results show that the constrained optimisation framework significantly improves system identification and prediction accuracy on the example of established state-of-the-art DSSMs. The EKVAE outperforms previous models w.r.t. prediction accuracy, achieves remarkable results in identifying dynamical systems, and can furthermore successfully learn state-space representations where static and dynamic features are disentangled.
****
On the Estimation Bias in Double Q-Learning
Zhizhou Ren, Guangxiang Zhu, Hao Hu, Beining Han, Jianglun Chen, Chongjie Zhang
Double Q-learning is a classical method for reducing overestimation bias, which is caused by taking maximum estimated values in the Bellman operation. Its variants in the deep Q-learning paradigm have shown great promise in producing reliable value prediction and improving learning performance. However, as shown by prior work, double Q-learning is not fully unbiased and suffers from underestimation bias. In this paper, we show that such underestimation bias may lead to multiple non-optimal fixed points under an approximate Bellman operator. To address the concerns of converging to non-optimal stationary solutions, we propose a simple but effective approach as a partial fix for the underestimation bias in double Q-learning. This approach leverages an approximate dynamic programming to bound the target value. We extensively evaluate our proposed method in the Atari benchmark tasks and demonstrate its significant improvement over baseline algorithms.
****
Mitigating Forgetting in Online Continual Learning with Neuron Calibration
Haiyan Yin, peng yang, Ping Li
Inspired by human intelligence, the research on online continual learning aims to push the limits of the machine learning models to constantly learn from sequentially encountered tasks, with the data from each task being observed in an online fashion. Though recent studies have achieved remarkable progress in improving the online continual learning performance empowered by the deep neural networks-based models, many of today's approaches still suffer a lot from catastrophic forgetting, a persistent challenge for continual learning. In this paper, we present a novel method which attempts to mitigate catastrophic forgetting in online continual learning from a new perspective, i.e., neuron calibration. In particular, we model the neurons in the deep neural networks-based models as calibrated units under a general formulation. Then we formalize a learning framework to effectively train the calibrated model, where neuron calibration could give ubiquitous benefit to balance the stability and plasticity of online continual learning algorithms through influencing both their forward inference path and backward optimization path. Our proposed formulation for neuron calibration is lightweight and applicable to general feed-forward neural networks-based models. We perform extensive experiments to evaluate our method on four benchmark continual learning datasets. The results show that neuron calibration plays a vital role in improving online continual learning performance and our method could substantially improve the state-of-the-art performance on all~the~evaluated~datasets.
****
Escaping Saddle Points with Compressed SGD
Dmitrii Avdiukhin, Grigory Yaroslavtsev
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****
Non-Gaussian Gaussian Processes for Few-Shot Regression
Marcin Sendera, Jacek Tabor, Aleksandra Nowak, Andrzej Bedychaj, Massimiliano Pa
tacchiola, Tomasz Trzcinski, Przemys■aw Spurek, Maciej Zieba
Gaussian Processes (GPs) have been widely used in machine learning to model dist
ributions over functions, with applications including multi-modal regression, ti
me-series prediction, and few-shot learning. GPs are particularly useful in the
last application since they rely on Normal distributions and enable closed-form
computation of the posterior probability function. Unfortunately, because the re
sulting posterior is not flexible enough to capture complex distributions, GPs a
ssume high similarity between subsequent tasks - a requirement rarely met in rea
l-world conditions. In this work, we address this limitation by leveraging the f
lexibility of Normalizing Flows to modulate the posterior predictive distributio
n of the GP. This makes the GP posterior locally non-Gaussian, therefore we name
our method Non-Gaussian Gaussian Processes (NGGPs). More precisely, we propose
an invertible ODE-based mapping that operates on each component of the random va
riable vectors and shares the parameters across all of them. We empirically test
ed the flexibility of NGGPs on various few-shot learning regression datasets, sh
owing that the mapping can incorporate context embedding information to model di
fferent noise levels for periodic functions. As a result, our method shares the
structure of the problem between subsequent tasks, but the contextualization all
ows for adaptation to dissimilarities. NGGPs outperform the competing state-of-t
he-art approaches on a diversified set of benchmarks and applications.
****
Believe What You See: Implicit Constraint Approach for Offline Multi-Agent Reinf
orcement Learning
Yiqin Yang, Xiaoteng Ma, Chenghao Li, Zewu Zheng, Qiyuan Zhang, Gao Huang, Jun Y
ang, Qianchuan Zhao
Learning from datasets without interaction with environments (Offline Learning)
is an essential step to apply Reinforcement Learning (RL) algorithms in real-wor
ld scenarios.However, compared with the single-agent counterpart, offline multi-
agent RL introduces more agents with the larger state and action space, which is
more challenging but attracts little attention. We demonstrate current offline
RL algorithms are ineffective in multi-agent systems due to the accumulated extr
apolation error. In this paper, we propose a novel offline RL algorithm, named I
mplicit Constraint Q-learning (ICQ), which effectively alleviates the extrapolat
ion error by only trusting the state-action pairs given in the dataset for value
estimation. Moreover, we extend ICQ to multi-agent tasks by decomposing the jo
int-policy under the implicit constraint. Experimental results demonstrate that
the extrapolation error is successfully controlled within a reasonable range an
d insensitive to the number of agents. We further show that ICQ achieves the sta
te-of-the-art performance in the challenging multi-agent offline tasks (StarCraf
t II). Our code is public online at https://github.com/YiqinYang/ICQ.
****
Online Learning in Periodic Zero-Sum Games
Tanner Fiez, Ryann Sim, Stratis Skoulakis, Georgios Piliouras, Lillian Ratliff
A seminal result in game theory is von Neumann's minmax theorem, which states th
at zero-sum games admit an essentially unique equilibrium solution. Classical le
arning results build on this theorem to show that online no-regret dynamics conv
erge to an equilibrium in a time-average sense in zero-sum games. In the past se
veral years, a key research direction has focused on characterizing the transien
t behavior of such dynamics. General results in this direction show that broad c
lasses of online learning dynamics are cyclic, and formally Poincar\'{e} recurre
nt, in zero-sum games. We analyze the robustness of these online learning behavi
ors in the case of periodic zero-sum games with a time-invariant equilibrium. Th
is model generalizes the usual repeated game formulation while also being a real
istic and natural model of a repeated competition between players that depends o
n exogenous environmental variations such as time-of-day effects, week-to-week t
rends, and seasonality. Interestingly, time-average convergence may fail even in
the simplest such settings, in spite of the equilibrium being fixed. In contras

t, using novel analysis methods, we show that Poincar\'{e} recurrence provably g eneralizes despite the complex, non-autonomous nature of these dynamical systems .
****

K-Net: Towards Unified Image Segmentation
Wenwei Zhang, Jiangmiao Pang, Kai Chen, Chen Change Loy
Semantic, instance, and panoptic segmentations have been addressed using differe nt and specialized frameworks despite their underlying connections. This paper p resents a unified, simple, and effective framework for these essentially similar tasks. The framework, named K-Net, segments both instances and semantic categor ies consistently by a group of learnable kernels, where each kernel is responsib le for generating a mask for either a potential instance or a stuff class. To re medy the difficulties of distinguishing various instances, we propose a kernel u pdate strategy that enables each kernel dynamic and conditional on its meaningfu l group in the input image. K-Net can be trained in an end-to-end manner with bi partite matching, and its training and inference are naturally NMS-free and box-free. Without bells and whistles, K-Net surpasses all previous published state-o f-the-art single-model results of panoptic segmentation on MS COCO test-dev spli t and semantic segmentation on ADE20K val split with 55.2% PQ and 54.3% mIoU, re spectively. Its instance segmentation performance is also on par with Cascade Ma sk R-CNN on MS COCO with 60%-90% faster inference speeds. Code and models will b e released at https://github.com/ZwwWayne/K-Net/.
****

Pareto-Optimal Learning-Augmented Algorithms for Online Conversion Problems
Bo Sun, Russell Lee, Mohammad Hajiesmaili, Adam Wierman, Danny Tsang
This paper leverages machine-learned predictions to design competitive algorithm s for online conversion problems with the goal of improving the competitive rati o when predictions are accurate (i.e., consistency), while also guaranteeing a w orst-case competitive ratio regardless of the prediction quality (i.e., robustne ss). We unify the algorithmic design of both integral and fractional conversion problems, which are also known as the 1-max-search and one-way trading problems, into a class of online threshold-based algorithms (OTA). By incorporating predi ctions into design of OTA, we achieve the Pareto-optimal trade-off of consistenc y and robustness, i.e., no online algorithm can achieve a better consistency gua rantee given for a robustness guarantee. We demonstrate the performance of OTA u sing numerical experiments on Bitcoin conversion.
****

Dynaboard: An Evaluation-As-A-Service Platform for Holistic Next-Generation Benc hmarking
Zhiyi Ma, Kawin Ethayarajh, Tristan Thrush, Somya Jain, Ledell Wu, Robin Jia, Ch ristopher Potts, Adina Williams, Douwe Kiela
We introduce Dynaboard, an evaluation-as-a-service framework for hosting benchma rks and conducting holistic model comparison, integrated with the Dynabench plat form. Our platform evaluates NLP models directly instead of relying on self-repo rted metrics or predictions on a single dataset. Under this paradigm, models are submitted to be evaluated in the cloud, circumventing the issues of reproducibi lity, accessibility, and backwards compatibility that often hinder benchmarking in NLP. This allows users to interact with uploaded models in real time to asses s their quality, and permits the collection of additional metrics such as memory use, throughput, and robustness, which -- despite their importance to practitio ners -- have traditionally been absent from leaderboards. On each task, models a re ranked according to the Dynascore, a novel utility-based aggregation of these statistics, which users can customize to better reflect their preferences, plac ing more/less weight on a particular axis of evaluation or dataset. As state-of-the-art NLP models push the limits of traditional benchmarks, Dynaboard offers a standardized solution for a more diverse and comprehensive evaluation of model quality.
****

NTopo: Mesh-free Topology Optimization using Implicit Neural Representations
Jonas Zehnder, Yue Li, Stelian Coros, Bernhard Thomaszewski

Recent advances in implicit neural representations show great promise when it comes to generating numerical solutions to partial differential equations. Compared to conventional alternatives, such representations employ parameterized neural networks to define, in a mesh-free manner, signals that are highly-detailed, continuous, and fully differentiable. In this work, we present a novel machine learning approach for topology optimization---an important class of inverse problems with high-dimensional parameter spaces and highly nonlinear objective landscapes. To effectively leverage neural representations in the context of mesh-free topology optimization, we use multilayer perceptrons to parameterize both density and displacement fields. Our experiments indicate that our method is highly competitive for minimizing structural compliance objectives, and it enables self-supervised learning of continuous solution spaces for topology optimization problems.
****
Generalization Bounds for (Wasserstein) Robust Optimization
Yang An, Rui Gao
(Distributionally) robust optimization has gained momentum in machine learning community recently, due to its promising applications in developing generalizable learning paradigms. In this paper, we derive generalization bounds for robust optimization and Wasserstein robust optimization for Lipschitz and piecewise Hölder smooth loss functions under both stochastic and adversarial setting, assuming that the underlying data distribution satisfies transportation-information inequalities. The proofs are built on new generalization bounds for variation regularization (such as Lipschitz or gradient regularization) and its connection with robustness.
****
Faster Matchings via Learned Duals
Michael Dinitz, Sungjin Im, Thomas Lavastida, Benjamin Moseley, Sergei Vassilvitskii
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 MDPs with linear function approximation and bandit feedback.

Gergely Neu, Julia Olkhovskaya
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 Collaborative Policies to Solve NP-hard Routing Problems
Minsu Kim, Jinkyoo Park, joungho kim
Recently, deep reinforcement learning (DRL) frameworks have shown potential for solving NP-hard routing problems such as the traveling salesman problem (TSP) without problem-specific expert knowledge. Although DRL can be used to solve complex problems, DRL frameworks still struggle to compete with state-of-the-art heuristics showing a substantial performance gap. This paper proposes a novel hierarchical problem-solving strategy, termed learning collaborative policies (LCP), which can effectively find the near-optimum solution using two iterative DRL policies: the seeder and reviser. The seeder generates as diversified candidate solutions as possible (seeds) while being dedicated to exploring over the full combinatorial action space (i.e., sequence of assignment action). To this end, we train the seeder's policy using a simple yet effective entropy regularization reward to encourage the seeder to find diverse solutions. On the other hand, the reviser modifies each candidate solution generated by the seeder; it partitions the full trajectory into sub-tours and simultaneously revises each sub-tour to minimize its traveling distance. Thus, the reviser is trained to improve the candidate solution's quality, focusing on the reduced solution space (which is beneficia

l for exploitation). Extensive experiments demonstrate that the proposed two-policies collaboration scheme improves over single-policy DRL framework on various NP-hard routing problems, including TSP, prize collecting TSP (PCTSP), and capacitated vehicle routing problem (CVRP).
****

Efficient Mirror Descent Ascent Methods for Nonsmooth Minimax Problems
Feihu Huang, Xidong Wu, Heng Huang
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

CO-PILOT: COllaborative Planning and reInforcement Learning On sub-Task curriculum
Shuang Ao, Tianyi Zhou, Guodong Long, Qinghua Lu, Liming Zhu, Jing Jiang
Goal-conditioned reinforcement learning (RL) usually suffers from sparse reward and inefficient exploration in long-horizon tasks. Planning can find the shortest path to a distant goal that provides dense reward/guidance but is inaccurate without a precise environment model. We show that RL and planning can collaboratively learn from each other to overcome their own drawbacks. In ''CO-PILOT'', a learnable path-planner and an RL agent produce dense feedback to train each other on a curriculum of tree-structured sub-tasks. Firstly, the planner recursively decomposes a long-horizon task to a tree of sub-tasks in a top-down manner, whose layers construct coarse-to-fine sub-task sequences as plans to complete the original task. The planning policy is trained to minimize the RL agent's cost of completing the sequence in each layer from top to bottom layers, which gradually increases the sub-tasks and thus forms an easy-to-hard curriculum for the planner. Next, a bottom-up traversal of the tree trains the RL agent from easier sub-tasks with denser rewards on bottom layers to harder ones on top layers and collects its cost on each sub-task train the planner in the next episode. CO-PILOT repeats this mutual training for multiple episodes before switching to a new task, so the RL agent and planner are fully optimized to facilitate each other's training. We compare CO-PILOT with RL (SAC, HER, PPO), planning (RRT*, NEXT, SGT), and their combination (SoRB) on navigation and continuous control tasks. CO-PILOT significantly improves the success rate and sample efficiency.
****

Modality-Agnostic Topology Aware Localization
Farhad Ghazvinian Zanjani, Ilia Karmanov, Hanno Ackermann, Daniel Dijkman, Simone Merlin, Max Welling, Fatih Porikli
This work presents a data-driven approach for the indoor localization of an observer on a 2D topological map of the environment. State-of-the-art techniques may yield accurate estimates only when they are tailor-made for a specific data modality like camera-based system that prevents their applicability to broader domains. Here, we establish a modality-agnostic framework (called OT-Isomap) and formulate the localization problem in the context of parametric manifold learning while leveraging optimal transportation. This framework allows jointly learning a low-dimensional embedding as well as correspondences with a topological map. We examine the generalizability of the proposed algorithm by applying it to data from diverse modalities such as image sequences and radio frequency signals. The experimental results demonstrate decimeter-level accuracy for localization using different sensory inputs.
****

Scalable Quasi-Bayesian Inference for Instrumental Variable Regression
Ziyu Wang, Yuhao Zhou, Tongzheng Ren, Jun Zhu
Recent years have witnessed an upsurge of interest in employing flexible machine learning models for instrumental variable (IV) regression, but the development of uncertainty quantification methodology is still lacking. In this work we present a scalable quasi-Bayesian procedure for IV regression, building upon the recently developed kernelized IV models. Contrary to Bayesian modeling for IV, our approach does not require additional assumptions on the data generating proces

s, and leads to a scalable approximate inference algorithm with time cost comparable to the corresponding point estimation methods. Our algorithm can be further extended to work with neural network models. We analyze the theoretical properties of the proposed quasi-posterior, and demonstrate through empirical evaluation the competitive performance of our method.
****
Kernel Identification Through Transformers
Fergus Simpson, Ian Davies, Vidhi Lalchand, Alessandro Vullo, Nicolas Durrande, Carl Edward Rasmussen
Kernel selection plays a central role in determining the performance of Gaussian Process (GP) models, as the chosen kernel determines both the inductive biases and prior support of functions under the GP prior. This work addresses the challenge of constructing custom kernel functions for high-dimensional GP regression models. Drawing inspiration from recent progress in deep learning, we introduce a novel approach named KITT: Kernel Identification Through Transformers. KITT exploits a transformer-based architecture to generate kernel recommendations in under 0.1 seconds, which is several orders of magnitude faster than conventional kernel search algorithms. We train our model using synthetic data generated from priors over a vocabulary of known kernels. By exploiting the nature of the self-attention mechanism, KITT is able to process datasets with inputs of arbitrary dimension. We demonstrate that kernels chosen by KITT yield strong performance over a diverse collection of regression benchmarks.
****
Curriculum Design for Teaching via Demonstrations: Theory and Applications
Gaurav Yengera, Rati Devidze, Parameswaran Kamalaruban, Adish Singla
We consider the problem of teaching via demonstrations in sequential decision-making settings. In particular, we study how to design a personalized curriculum over demonstrations to speed up the learner's convergence. We provide a unified curriculum strategy for two popular learner models: Maximum Causal Entropy Inverse Reinforcement Learning (MaxEnt-IRL) and Cross-Entropy Behavioral Cloning (CrossEnt-BC). Our unified strategy induces a ranking over demonstrations based on a notion of difficulty scores computed w.r.t. the teacher's optimal policy and the learner's current policy. Compared to the state of the art, our strategy doesn't require access to the learner's internal dynamics and still enjoys similar convergence guarantees under mild technical conditions. Furthermore, we adapt our curriculum strategy to the setting where no teacher agent is present using task-specific difficulty scores. Experiments on a synthetic car driving environment and navigation-based environments demonstrate the effectiveness of our curriculum strategy.
****
Revenue maximization via machine learning with noisy data
Ellen Vitercik, Tom Yan
Increasingly, copious amounts of consumer data are used to learn high-revenue mechanisms via machine learning. Existing research on mechanism design via machine learning assumes that there is a distribution over the buyers' values for the items for sale and that the learning algorithm's input is a training set sampled from this distribution. This setup makes the strong assumption that no noise is introduced during data collection. In order to help place mechanism design via machine learning on firm foundations, we investigate the extent to which this learning process is robust to noise. Optimizing revenue using noisy data is challenging because revenue functions are extremely volatile: an infinitesimal change in the buyers' values can cause a steep drop in revenue. Nonetheless, we provide guarantees when arbitrarily correlated noise is added to the training set; we only require that the noise has bounded magnitude or is sub-Gaussian. We conclude with an application of our guarantees to multi-task mechanism design, where there are multiple distributions over buyers' values and the goal is to learn a high-revenue mechanism per distribution. To our knowledge, we are the first to study mechanism design via machine learning with noisy data as well as multi-task mechanism design.
****

Exploiting Data Sparsity in Secure Cross-Platform Social Recommendation
Jinming Cui, Chaochao Chen, Lingjuan Lyu, Carl Yang, Wang Li
Social recommendation has shown promising improvements over traditional systems since it leverages social correlation data as an additional input. Most existing work assumes that all data are available to the recommendation platform. However, in practice, user-item interaction data (e.g.,rating) and user-user social data are usually generated by different platforms, and both of which contain sensitive information. Therefore, "How to perform secure and efficient social recommendation across different platforms, where the data are highly-sparse in nature" remains an important challenge. In this work, we bring secure computation techniques into social recommendation, and propose S3Rec, a sparsity-aware secure cross-platform social recommendation framework. As a result, our model can not only improve the recommendation performance of the rating platform by incorporating the sparse social data on the social platform, but also protect data privacy of both platforms. Moreover, to further improve model training efficiency, we propose two secure sparse matrix multiplication protocols based on homomorphic encryption and private information retrieval. Our experiments on two benchmark datasets demonstrate the effectiveness of S3Rec.
****
Parallelizing Thompson Sampling
Amin Karbasi, Vahab Mirrokni, Mohammad Shadravan
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Dynamic Causal Bayesian Optimization
Virginia Aglietti, Neil Dhir, Javier González, Theodoros Damoulas
We study the problem of performing a sequence of optimal interventions in a dynamic causal system where both the target variable of interest, and the inputs, evolve over time. This problem arises in a variety of domains including healthcare, operational research and policy design. Our approach, which we call Dynamic Causal Bayesian Optimisation (DCBO), brings together ideas from decision making, causal inference and Gaussian process (GP) emulation. DCBO is useful in scenarios where the causal effects are changing over time. Indeed, at every time step, DCBO identifies a local optimal intervention by integrating both observational and past interventional data collected from the system. We give theoretical results detailing how one can transfer interventional information across time steps and define a dynamic causal GP model which can be used to find optimal interventions in practice. Finally, we demonstrate how DCBO identifies optimal interventions faster than competing approaches in multiple settings and applications.
****
Local Differential Privacy for Regret Minimization in Reinforcement Learning
Evrard Garcelon, Vianney Perchet, Ciara Pike-Burke, Matteo Pirotta
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Emergent Discrete Communication in Semantic Spaces
Mycal Tucker, Huao Li, Siddharth Agrawal, Dana Hughes, Katia Sycara, Michael Lewis, Julie A Shah
Neural agents trained in reinforcement learning settings can learn to communicate among themselves via discrete tokens, accomplishing as a team what agents would be unable to do alone. However, the current standard of using one-hot vectors as discrete communication tokens prevents agents from acquiring more desirable aspects of communication such as zero-shot understanding. Inspired by word embedding techniques from natural language processing, we propose neural agent architectures that enables them to communicate via discrete tokens derived from a learned, continuous space. We show in a decision theoretic framework that our techniq

ue optimizes communication over a wide range of scenarios, whereas one-hot tokens are only optimal under restrictive assumptions. In self-play experiments, we validate that our trained agents learn to cluster tokens in semantically-meaningful ways, allowing them communicate in noisy environments where other techniques fail. Lastly, we demonstrate both that agents using our method can effectively respond to novel human communication and that humans can understand unlabeled emergent agent communication, outperforming the use of one-hot communication.
****
Drop, Swap, and Generate: A Self-Supervised Approach for Generating Neural Activity
Ran Liu, Mehdi Azabou, Max Dabagia, Chi-Heng Lin, Mohammad Gheshlaghi Azar, Keith Hengen, Michal Valko, Eva Dyer
Meaningful and simplified representations of neural activity can yield insights into how and what information is being processed within a neural circuit. However, without labels, finding representations that reveal the link between the brain and behavior can be challenging. Here, we introduce a novel unsupervised approach for learning disentangled representations of neural activity called Swap-VAE. Our approach combines a generative modeling framework with an instance-specific alignment loss that tries to maximize the representational similarity between transformed views of the input (brain state). These transformed (or augmented) views are created by dropping out neurons and jittering samples in time, which intuitively should lead the network to a representation that maintains both temporal consistency and invariance to the specific neurons used to represent the neural state. Through evaluations on both synthetic data and neural recordings from hundreds of neurons in different primate brains, we show that it is possible to build representations that disentangle neural datasets along relevant latent dimensions linked to behavior.
****
Equivariant Manifold Flows
Isay Katsman, Aaron Lou, Derek Lim, Qingxuan Jiang, Ser Nam Lim, Christopher M. De Sa
Tractably modelling distributions over manifolds has long been an important goal in the natural sciences. Recent work has focused on developing general machine learning models to learn such distributions. However, for many applications these distributions must respect manifold symmetries—a trait which most previous models disregard. In this paper, we lay the theoretical foundations for learning symmetry-invariant distributions on arbitrary manifolds via equivariant manifold flows. We demonstrate the utility of our approach by learning quantum field theory-motivated invariant SU(n) densities and by correcting meteor impact dataset bias.
****
Scalable Bayesian GPFA with automatic relevance determination and discrete noise models
Kristopher Jensen, Ta-Chu Kao, Jasmine Stone, Guillaume Hennequin
Latent variable models are ubiquitous in the exploratory analysis of neural population recordings, where they allow researchers to summarize the activity of large populations of neurons in lower dimensional ‘latent’ spaces. Existing methods can generally be categorized into (i) Bayesian methods that facilitate flexible incorporation of prior knowledge and uncertainty estimation, but which typically do not scale to large datasets; and (ii) highly parameterized methods without explicit priors that scale better but often struggle in the low-data regime. Here, we bridge this gap by developing a fully Bayesian yet scalable version of Gaussian process factor analysis (bGPFA), which models neural data as arising from a set of inferred latent processes with a prior that encourages smoothness over time. Additionally, bGPFA uses automatic relevance determination to infer the dimensionality of neural activity directly from the training data during optimization. To enable the analysis of continuous recordings without trial structure, we introduce a novel variational inference strategy that scales near-linearly in time and also allows for non-Gaussian noise models appropriate for electrophysiological recordings. We apply bGPFA to continuous recordings spanning 30 minutes w

ith over 14 million data points from primate motor and somatosensory cortices during a self-paced reaching task. We show that neural activity progresses from an initial state at target onset to a reach- specific preparatory state well before movement onset. The distance between these initial and preparatory latent states is predictive of reaction times across reaches, suggesting that such preparatory dynamics have behavioral relevance despite the lack of externally imposed delay periods. Additionally, bGPFA discovers latent processes that evolve over slow timescales on the order of several seconds and contain complementary information about reaction time. These timescales are longer than those revealed by methods which focus on individual movement epochs and may reflect fluctuations in e.g. task engagement.
****
Recurrence along Depth: Deep Convolutional Neural Networks with Recurrent Layer Aggregation
Jingyu Zhao, Yanwen Fang, Guodong Li
This paper introduces a concept of layer aggregation to describe how information from previous layers can be reused to better extract features at the current layer. While DenseNet is a typical example of the layer aggregation mechanism, its redundancy has been commonly criticized in the literature. This motivates us to propose a very light-weighted module, called recurrent layer aggregation (RLA), by making use of the sequential structure of layers in a deep CNN. Our RLA module is compatible with many mainstream deep CNNs, including ResNets, Xception and MobileNetV2, and its effectiveness is verified by our extensive experiments on image classification, object detection and instance segmentation tasks. Specifically, improvements can be uniformly observed on CIFAR, ImageNet and MS COCO datasets, and the corresponding RLA-Nets can surprisingly boost the performances by 2-3% on the object detection task. This evidences the power of our RLA module in helping main CNNs better learn structural information in images.
****
Independent Prototype Propagation for Zero-Shot Compositionality
Frank Ruis, Gertjan Burghouts, Doina Bucur
Humans are good at compositional zero-shot reasoning; someone who has never seen a zebra before could nevertheless recognize one when we tell them it looks like a horse with black and white stripes. Machine learning systems, on the other hand, usually leverage spurious correlations in the training data, and while such correlations can help recognize objects in context, they hurt generalization. To be able to deal with underspecified datasets while still leveraging contextual clues during classification, we propose ProtoProp, a novel prototype propagation graph method. First we learn prototypical representations of objects (e.g., zebra) that are independent w.r.t. their attribute labels (e.g., stripes) and vice versa. Next we propagate the independent prototypes through a compositional graph, to learn compositional prototypes of novel attribute-object combinations that reflect the dependencies of the target distribution. The method does not rely on any external data, such as class hierarchy graphs or pretrained word embeddings. We evaluate our approach on AO-Clevr, a synthetic and strongly visual dataset with clean labels, UT-Zappos, a noisy real-world dataset of fine-grained shoe types, and C-GQA, a large-scale object detection dataset modified for compositional zero-shot learning. We show that in the generalized compositional zero-shot setting we outperform state-of-the-art results, and through ablations we show the importance of each part of the method and their contribution to the final results. The code is available on github.
****
Universal Graph Convolutional Networks
Di Jin, Zhizhi Yu, Cuiying Huo, Rui Wang, Xiao Wang, Dongxiao He, Jiawei Han
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Adversarial Feature Desensitization

Pouya Bashivan, Reza Bayat, Adam Ibrahim, Kartik Ahuja, Mojtaba Faramarzi, Touraj Laleh, Blake Richards, Irina Rish
Neural networks are known to be vulnerable to adversarial attacks -- slight but carefully constructed perturbations of the inputs which can drastically impair the network's performance. Many defense methods have been proposed for improving robustness of deep networks by training them on adversarially perturbed inputs. However, these models often remain vulnerable to new types of attacks not seen during training, and even to slightly stronger versions of previously seen attacks. In this work, we propose a novel approach to adversarial robustness, which builds upon the insights from the domain adaptation field. Our method, called Adversarial Feature Desensitization (AFD), aims at learning features that are invariant towards adversarial perturbations of the inputs. This is achieved through a game where we learn features that are both predictive and robust (insensitive to adversarial attacks), i.e. cannot be used to discriminate between natural and adversarial data. Empirical results on several benchmarks demonstrate the effectiveness of the proposed approach against a wide range of attack types and attack strengths. Our code is available at https://github.com/BashivanLab/afd.
****
Few-Shot Data-Driven Algorithms for Low Rank Approximation
Piotr Indyk, Tal Wagner, David Woodruff
Recently, data-driven and learning-based algorithms for low rank matrix approximation were shown to outperform classical data-oblivious algorithms by wide margins in terms of accuracy. Those algorithms are based on the optimization of sparse sketching matrices, which lead to large savings in time and memory during testing. However, they require long training times on a large amount of existing data, and rely on access to specialized hardware and software. In this work, we develop new data-driven low rank approximation algorithms with better computational efficiency in the training phase, alleviating these drawbacks. Furthermore, our methods are interpretable: while previous algorithms choose the sketching matrix either at random or by black-box learning, we show that it can be set (or initialized) to clearly interpretable values extracted from the dataset. Our experiments show that our algorithms, either by themselves or in combination with previous methods, achieve significant empirical advantage over previous work, improving training times by up to an order of magnitude toward achieving the same target accuracy.
****
Neural-PIL: Neural Pre-Integrated Lighting for Reflectance Decomposition
Mark Boss, Varun Jampani, Raphael Braun, Ce Liu, Jonathan Barron, Hendrik PA Lensch
Decomposing a scene into its shape, reflectance and illumination is a fundamental problem in computer vision and graphics. Neural approaches such as NeRF have achieved remarkable success in view synthesis, but do not explicitly perform decomposition and instead operate exclusively on radiance (the product of reflectance and illumination). Extensions to NeRF, such as NeRD, can perform decomposition but struggle to accurately recover detailed illumination, thereby significantly limiting realism. We propose a novel reflectance decomposition network that can estimate shape, BRDF, and per-image illumination given a set of object images captured under varying illumination. Our key technique is a novel illumination integration network called Neural-PIL that replaces a costly illumination integral operation in the rendering with a simple network query. In addition, we also learn deep low-dimensional priors on BRDF and illumination representations using novel smooth manifold auto-encoders. Our decompositions can result in considerably better BRDF and light estimates enabling more accurate novel view-synthesis and relighting compared to prior art. Project page: https://markboss.me/publication/2021-neural-pil/
****
Asymptotics of the Bootstrap via Stability with Applications to Inference with Model Selection
Morgane Austern, Vasilis Syrgkanis
One of the most commonly used methods for forming confidence intervals is the em

pirical bootstrap, which is especially expedient when the limiting distribution of the estimator is unknown. However, despite its ubiquitous role in machine learning, its theoretical properties are still not well understood. Recent developments in probability have provided new tools to study the bootstrap method. However, they have been applied only to specific applications and contexts, and it is unclear whether these techniques are applicable to the understanding of the consistency of the bootstrap in machine learning pipelines. In this paper, we derive general stability conditions under which the empirical bootstrap estimator is consistent and quantify the speed of convergence. Moreover, we propose alternative ways to use the bootstrap method to build confidence intervals with coverage guarantees. Finally, we illustrate the generality and tightness of our results by examples of interest for machine learning including for two-sample kernel tests after kernel selection and the empirical risk of stacked estimators.
****
Dynamic influence maximization
Binghui 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-authors prior to requesting a name change in the electronic proceedings.
****
Risk Monotonicity in Statistical Learning
Zakaria Mhammedi
Acquisition of data is a difficult task in many applications of machine learning, and it is only natural that one hopes and expects the population risk to decrease (better performance) monotonically with increasing data points. It turns out, somewhat surprisingly, that this is not the case even for the most standard algorithms that minimize the empirical risk. Non-monotonic behavior of the risk and instability in training have manifested and appeared in the popular deep learning paradigm under the description of double descent. These problems highlight the current lack of understanding of learning algorithms and generalization. It is, therefore, crucial to pursue this concern and provide a characterization of such behavior. In this paper, we derive the first consistent and risk-monotonic (in high probability) algorithms for a general statistical learning setting under weak assumptions, consequently answering some questions posed by Viering et. al. 2019 on how to avoid non-monotonic behavior of risk curves. We further show that risk monotonicity need not necessarily come at the price of worse excess risk rates. To achieve this, we derive new empirical Bernstein-like concentration inequalities of independent interest that hold for certain non-i.i.d.~processes such as Martingale Difference Sequences.
****
Information is Power: Intrinsic Control via Information Capture
Nicholas Rhinehart, Jenny Wang, Glen Berseth, John Co-Reyes, Danijar Hafner, Chelsea Finn, Sergey Levine
Humans and animals explore their environment and acquire useful skills even in the absence of clear goals, exhibiting intrinsic motivation. The study of intrinsic motivation in artificial agents is concerned with the following question: what is a good general-purpose objective for an agent? We study this question in dynamic partially-observed environments, and argue that a compact and general learning objective is to minimize the entropy of the agent's state visitation estimated using a latent state-space model. This objective induces an agent to both gather information about its environment, corresponding to reducing uncertainty, and to gain control over its environment, corresponding to reducing the unpredictability of future world states. We instantiate this approach as a deep reinforcement learning agent equipped with a deep variational Bayes filter. We find that our agent learns to discover, represent, and exercise control of dynamic objects in a variety of partially-observed environments sensed with visual observations without extrinsic reward.
****
Extracting Deformation-Aware Local Features by Learning to Deform

Guilherme Potje, Renato Martins, Felipe Chamone, Erickson Nascimento
Despite the advances in extracting local features achieved by handcrafted and le
arning-based descriptors, they are still limited by the lack of invariance to no
n-rigid transformations. In this paper, we present a new approach to compute fea
tures from still images that are robust to non-rigid deformations to circumvent
the problem of matching deformable surfaces and objects. Our deformation-aware l
ocal descriptor, named DEAL, leverages a polar sampling and a spatial transforme
r warping to provide invariance to rotation, scale, and image deformations. We t
rain the model architecture end-to-end by applying isometric non-rigid deformati
ons to objects in a simulated environment as guidance to provide highly discrimi
native local features. The experiments show that our method outperforms state-of
-the-art handcrafted, learning-based image, and RGB-D descriptors in different d
atasets with both real and realistic synthetic deformable objects in still image
s. The source code and trained model of the descriptor are publicly available at
 https://www.verlab.dcc.ufmg.br/descriptors/neurips2021.
****
Object-Centric Representation Learning with Generative Spatial-Temporal Factoriz
ation
Nanbo Li, Muhammad Ahmed Raza, Wenbin Hu, Zhaole Sun, Robert Fisher
Learning object-centric scene representations is essential for attaining structu
ral understanding and abstraction of complex scenes. Yet, as current approaches
for unsupervised object-centric representation learning are built upon either a
stationary observer assumption or a static scene assumption, they often: i) suff
er single-view spatial ambiguities, or ii) infer incorrectly or inaccurately obj
ect representations from dynamic scenes. To address this, we propose Dynamics-aw
are Multi-Object Network (DyMON), a method that broadens the scope of multi-view
 object-centric representation learning to dynamic scenes. We train DyMON on mul
ti-view-dynamic-scene data and show that DyMON learns---without supervision---to
 factorize the entangled effects of observer motions and scene object dynamics f
rom a sequence of observations, and constructs scene object spatial representati
ons suitable for rendering at arbitrary times (querying across time) and from ar
bitrary viewpoints (querying across space). We also show that the factorized sce
ne representations (w.r.t. objects) support querying about a single object by sp
ace and time independently.
****
Learning to Simulate Self-driven Particles System with Coordinated Policy Optimi
zation
Zhenghao Peng, Quanyi Li, Ka Ming Hui, Chunxiao Liu, Bolei Zhou
Self-Driven Particles (SDP) describe a category of multi-agent systems common in
 everyday life, such as flocking birds and traffic flows. In a SDP system, each
agent pursues its own goal and constantly changes its cooperative or competitive
 behaviors with its nearby agents. Manually designing the controllers for such S
DP system is time-consuming, while the resulting emergent behaviors are often no
t realistic nor generalizable. Thus the realistic simulation of SDP systems rema
ins challenging. Reinforcement learning provides an appealing alternative for au
tomating the development of the controller for SDP. However, previous multi-agen
t reinforcement learning (MARL) methods define the agents to be teammates or ene
mies before hand, which fail to capture the essence of SDP where the role of eac
h agent varies to be cooperative or competitive even within one episode. To simu
late SDP with MARL, a key challenge is to coordinate agents' behaviors while sti
ll maximizing individual objectives. Taking traffic simulation as the testing be
d, in this work we develop a novel MARL method called Coordinated Policy Optimiz
ation (CoPO), which incorporates social psychology principle to learn neural con
troller for SDP. Experiments show that the proposed method can achieve superior
performance compared to MARL baselines in various metrics. Noticeably the traine
d vehicles exhibit complex and diverse social behaviors that improve performance
 and safety of the population as a whole. Demo video and source code are availab
le at: https://decisionforce.github.io/CoPO/
****
Gradient-based Hyperparameter Optimization Over Long Horizons

Paul Micaelli, Amos J. Storkey
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Stochastic Bias-Reduced Gradient Methods
Hilal Asi, Yair Carmon, Arun Jambulapati, Yujia Jin, Aaron Sidford
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
The Causal-Neural Connection: Expressiveness, Learnability, and Inference
Kevin Xia, Kai-Zhan Lee, Yoshua Bengio, Elias Bareinboim
One of the central elements of any causal inference is an object called structural causal model (SCM), which represents a collection of mechanisms and exogenous sources of random variation of the system under investigation (Pearl, 2000). An important property of many kinds of neural networks is universal approximability: the ability to approximate any function to arbitrary precision. Given this property, one may be tempted to surmise that a collection of neural nets is capable of learning any SCM by training on data generated by that SCM. In this paper, we show this is not the case by disentangling the notions of expressivity and learnability. Specifically, we show that the causal hierarchy theorem (Thm. 1, Bareinboim et al., 2020), which describes the limits of what can be learned from data, still holds for neural models. For instance, an arbitrarily complex and expressive neural net is unable to predict the effects of interventions given observational data alone. Given this result, we introduce a special type of SCM called a neural causal model (NCM), and formalize a new type of inductive bias to encode structural constraints necessary for performing causal inferences. Building on this new class of models, we focus on solving two canonical tasks found in the literature known as causal identification and estimation. Leveraging the neural toolbox, we develop an algorithm that is both sufficient and necessary to determine whether a causal effect can be learned from data (i.e., causal identifiability); it then estimates the effect whenever identifiability holds (causal estimation). Simulations corroborate the proposed approach.
****
Validation Free and Replication Robust Volume-based Data Valuation
Xinyi Xu, Zhaoxuan Wu, Chuan Sheng Foo, Bryan Kian Hsiang Low
Data valuation arises as a non-trivial challenge in real-world use cases such as collaborative machine learning, federated learning, trusted data sharing, data marketplaces. The value of data is often associated with the learning performance (e.g., validation accuracy) of a model trained on the data, which introduces a close coupling between data valuation and validation. However, a validation set may notbe available in practice and it can be challenging for the data providers to reach an agreement on the choice of the validation set. Another practical issue is that of data replication: Given the value of some data points, a dishonest data provider may replicate these data points to exploit the valuation for a larger reward/payment. We observe that the diversity of the data points is an inherent property of a dataset that is independent of validation. We formalize diversity via the volume of the data matrix (i.e., determinant of its left Gram), which allows us to establish a formal connection between the diversity of data and learning performance without requiring validation. Furthermore, we propose a robust volume measure with a theoretical guarantee on the replication robustness by following the intuition that copying the same data points does not increase the diversity of data. We perform extensive experiments to demonstrate its consistency in valuation and practical advantages over existing baselines and show that our method is model- and task-agnostic and can be flexibly adapted to handle various neural networks.
****

Implicit Finite-Horizon Approximation and Efficient Optimal Algorithms for Stochastic Shortest Path
Liyu Chen, Mehdi Jafarnia-Jahromi, Rahul Jain, Haipeng Luo
We introduce a generic template for developing regret minimization algorithms in the Stochastic Shortest Path (SSP) model, which achieves minimax optimal regret as long as certain properties are ensured. The key of our analysis is a new technique called implicit finite-horizon approximation, which approximates the SSP model by a finite-horizon counterpart only in the analysis without explicit implementation. Using this template, we develop two new algorithms: the first one is model-free (the first in the literature to our knowledge) and minimax optimal under strictly positive costs; the second one is model-based and minimax optimal even with zero-cost state-action pairs, matching the best existing result from [Tarbouriech et al., 2021b]. Importantly, both algorithms admit highly sparse updates, making them computationally more efficient than all existing algorithms. Moreover, both can be made completely parameter-free.
****
A Separation Result Between Data-oblivious and Data-aware Poisoning Attacks
Samuel Deng, Sanjam Garg, Somesh Jha, Saeed Mahloujifar, Mohammad Mahmoody, Abhradeep Guha Thakurta
Poisoning attacks have emerged as a significant security threat to machine learning algorithms. It has been demonstrated that adversaries who make small changes to the training set, such as adding specially crafted data points, can hurt the performance of the output model. Most of these attacks require the full knowledge of training data. This leaves open the possibility of achieving the same attack results using poisoning attacks that do not have the full knowledge of the clean training set.In this work, we initiate a theoretical study of the problem above. Specifically, for the case of feature selection with LASSO, we show that \emph{full information} adversaries (that craft poisoning examples based on the rest of the training data) are provably much more devastating compared to the optimal attacker that is \emph{oblivious} to the training set yet has access to the distribution of the data. Our separation result shows that the two settings of data-aware and data-oblivious are fundamentally different and we cannot hope to achieve the same attack or defense results in these scenarios.
****
Deep Learning Through the Lens of Example Difficulty
Robert Baldock, Hartmut Maennel, Behnam Neyshabur
Existing work on understanding deep learning often employs measures that compress all data-dependent information into a few numbers. In this work, we adopt a perspective based on the role of individual examples. We introduce a measure of the computational difficulty of making a prediction for a given input: the (effective) prediction depth. Our extensive investigation reveals surprising yet simple relationships between the prediction depth of a given input and the model’s uncertainty, confidence, accuracy and speed of learning for that data point. We further categorize difficult examples into three interpretable groups, demonstrate how these groups are processed differently inside deep models and showcase how this understanding allows us to improve prediction accuracy. Insights from our study lead to a coherent view of a number of separately reported phenomena in the literature: early layers generalize while later layers memorize; early layers converge faster and networks learn easy data and simple functions first.
****
R-Drop: Regularized Dropout for Neural Networks
xiaobo liang, Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, Tie-Yan Liu
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Diversity Enhanced Active Learning with Strictly Proper Scoring Rules
Wei Tan, Lan Du, Wray Buntine

We study acquisition functions for active learning (AL) for text classification. The Expected Loss Reduction (ELR) method focuses on a Bayesian estimate of the reduction in classification error, recently updated with Mean Objective Cost of Uncertainty (MOCU). We convert the ELR framework to estimate the increase in (strictly proper) scores like log probability or negative mean square error, which we call Bayesian Estimate of Mean Proper Scores (BEMPS). We also prove convergence results borrowing techniques used with MOCU. In order to allow better experimentation with the new acquisition functions, we develop a complementary batch AL algorithm, which encourages diversity in the vector of expected changes in scores for unlabelled data. To allow high performance text classifiers, we combine ensembling and dynamic validation set construction on pretrained language models. Extensive experimental evaluation then explores how these different acquisition functions perform. The results show that the use of mean square error and log probability with BEMPS yields robust acquisition functions, which consistently outperform the others tested.
****
SSUL: Semantic Segmentation with Unknown Label for Exemplar-based Class-Incremental Learning
Sungmin Cha, beomyoung kim, YoungJoon Yoo, Taesup Moon
We consider a class-incremental semantic segmentation (CISS) problem. While some recently proposed algorithms utilized variants of knowledge distillation (KD) technique to tackle the problem, they only partially addressed the key additional challenges in CISS that causes the catastrophic forgetting; \textit{i.e.}, the semantic drift of the background class and multi-label prediction issue. To better address these challenges, we propose a new method, dubbed as SSUL-M (Semantic Segmentation with Unknown Label with Memory), by carefully combining several techniques tailored for semantic segmentation. More specifically, we make three main contributions; (1) modeling \textit{unknown} class within the background class to help learning future classes (help plasticity), (2) \textit{freezing} backbone network and past classifiers with binary cross-entropy loss and pseudo-labeling to overcome catastrophic forgetting (help stability), and (3) utilizing \textit{tiny exemplar memory} for the first time in CISS to improve \textit{both} plasticity and stability. As a result, we show our method achieves significantly better performance than the recent state-of-the-art baselines on the standard benchmark datasets. Furthermore, we justify our contributions with thorough and extensive ablation analyses and discuss different natures of the CISS problem compared to the standard class-incremental learning for classification. The official code is available at https://github.com/clovaai/SSUL.
****
Lower and Upper Bounds on the Pseudo-Dimension of Tensor Network Models
Behnoush Khavari, Guillaume Rabusseau
Tensor network methods have been a key ingredient of advances in condensed matter physics and have recently sparked interest in the machine learning community for their ability to compactly represent very high-dimensional objects. Tensor network methods can for example be used to efficiently learn linear models in exponentially large feature spaces [Stoudenmire and Schwab, 2016]. In this work, we derive upper and lower bounds on the VC dimension and pseudo-dimension of a large class of tensor network models for classification, regression and completion. Our upper bounds hold for linear models parameterized by arbitrary tensor network structures, and we derive lower bounds for common tensor decomposition models ~(CP, Tensor Train, Tensor Ring and Tucker) showing the tightness of our general upper bound. These results are used to derive a generalization bound which can be applied to classification with low rank matrices as well as linear classifiers based on any of the commonly used tensor decomposition models. As a corollary of our results, we obtain a bound on the VC dimension of the matrix product state classifier introduced in [Stoudenmire and Schwab, 2016] as a function of the so-called bond dimension~(i.e. tensor train rank), which answers an open problem listed by Cirac, Garre-Rubio and Pérez-García in [Cirac et al., 2019].
****
What Makes Multi-Modal Learning Better than Single (Provably)

Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, Longbo Huang

The world provides us with data of multiple modalities. Intuitively, models fusing data from different modalities outperform their uni-modal counterparts, since more information is aggregated. Recently, joining the success of deep learning, there is an influential line of work on deep multi-modal learning, which has remarkable empirical results on various applications. However, theoretical justifications in this field are notably lacking. Can multi-modal learning provably perform better than uni-modal?In this paper, we answer this question under a most popular multi-modal fusion framework, which firstly encodes features from different modalities into a common latent space and seamlessly maps the latent representations into the task space. We prove that learning with multiple modalities achieves a smaller population risk than only using its subset of modalities. The main intuition is that the former has a more accurate estimate of the latent space representation. To the best of our knowledge, this is the first theoretical treatment to capture important qualitative phenomena observed in real multi-modal applications from the generalization perspective. Combining with experiment results, we show that multi-modal learning does possess an appealing formal guarantee.

****

Quantifying and Improving Transferability in Domain Generalization

Guojun Zhang, Han Zhao, Yaoliang Yu, Pascal Poupart

Out-of-distribution generalization is one of the key challenges when transferring a model from the lab to the real world. Existing efforts mostly focus on building invariant features among source and target domains. Based on invariant features, a high-performing classifier on source domains could hopefully behave equally well on a target domain. In other words, we hope the invariant features to be \emph{transferable}. However, in practice, there are no perfectly transferable features, and some algorithms seem to learn ``more transferable'' features than others. How can we understand and quantify such \emph{transferability}? In this paper, we formally define transferability that one can quantify and compute in domain generalization. We point out the difference and connection with common discrepancy measures between domains, such as total variation and Wasserstein distance. We then prove that our transferability can be estimated with enough samples and give a new upper bound for the target error based on our transferability. Empirically, we evaluate the transferability of the feature embeddings learned by existing algorithms for domain generalization. Surprisingly, we find that many algorithms are not quite learning transferable features, although few could still survive. In light of this, we propose a new algorithm for learning transferable features and test it over various benchmark datasets, including RotatedMNIST, PACS, Office-Home and WILDS-FMoW. Experimental results show that the proposed algorithm achieves consistent improvement over many state-of-the-art algorithms, corroborating our theoretical findings.

****

Beyond Pinball Loss: Quantile Methods for Calibrated Uncertainty Quantification

Youngseog Chung, Willie Neiswanger, Ian Char, Jeff Schneider

Among the many ways of quantifying uncertainty in a regression setting, specifying the full quantile function is attractive, as quantiles are amenable to interpretation and evaluation. A model that predicts the true conditional quantiles for each input, at all quantile levels, presents a correct and efficient representation of the underlying uncertainty. To achieve this, many current quantile-based methods focus on optimizing the pinball loss. However, this loss restricts the scope of applicable regression models, limits the ability to target many desirable properties (e.g. calibration, sharpness, centered intervals), and may produce poor conditional quantiles. In this work, we develop new quantile methods that address these shortcomings. In particular, we propose methods that can apply to any class of regression model, select an explicit balance between calibration and sharpness, optimize for calibration of centered intervals, and produce more accurate conditional quantiles. We provide a thorough experimental evaluation of our methods, which includes a high dimensional uncertainty quantification task in nuclear fusion.

****
Dynamic Inference with Neural Interpreters
Nasim Rahaman, Muhammad Waleed Gondal, Shruti Joshi, Peter Gehler, Yoshua Bengio, Francesco Locatello, Bernhard Schölkopf
Modern neural network architectures can leverage large amounts of data to generalize well within the training distribution. However, they are less capable of systematic generalization to data drawn from unseen but related distributions, a feat that is hypothesized to require compositional reasoning and reuse of knowledge. In this work, we present Neural Interpreters, an architecture that factorizes inference in a self-attention network as a system of modules, which we call functions. Inputs to the model are routed through a sequence of functions in a way that is end-to-end learned. The proposed architecture can flexibly compose computation along width and depth, and lends itself well to capacity extension after training. To demonstrate the versatility of Neural Interpreters, we evaluate it in two distinct settings: image classification and visual abstract reasoning on Raven Progressive Matrices. In the former, we show that Neural Interpreters perform on par with the vision transformer using fewer parameters, while being transferrable to a new task in a sample efficient manner. In the latter, we find that Neural Interpreters are competitive with respect to the state-of-the-art in terms of systematic generalization.
****
Leveraging Recursive Gumbel-Max Trick for Approximate Inference in Combinatorial Spaces
Kirill Struminsky, Artyom Gadetsky, Denis Rakitin, Danil Karpushkin, Dmitry P. Vetrov
Structured latent variables allow incorporating meaningful prior knowledge into deep learning models. However, learning with such variables remains challenging because of their discrete nature. Nowadays, the standard learning approach is to define a latent variable as a perturbed algorithm output and to use a differentiable surrogate for training. In general, the surrogate puts additional constraints on the model and inevitably leads to biased gradients. To alleviate these shortcomings, we extend the Gumbel-Max trick to define distributions over structured domains. We avoid the differentiable surrogates by leveraging the score function estimators for optimization. In particular, we highlight a family of recursive algorithms with a common feature we call stochastic invariant. The feature allows us to construct reliable gradient estimates and control variates without additional constraints on the model. In our experiments, we consider various structured latent variable models and achieve results competitive with relaxation-based counterparts.
****
Hamiltonian Dynamics with Non-Newtonian Momentum for Rapid Sampling
Greg Ver Steeg, Aram Galstyan
Sampling from an unnormalized probability distribution is a fundamental problem in machine learning with applications including Bayesian modeling, latent factor inference, and energy-based model training. After decades of research, variations of MCMC remain the default approach to sampling despite slow convergence. Auxiliary neural models can learn to speed up MCMC, but the overhead for training the extra model can be prohibitive. We propose a fundamentally different approach to this problem via a new Hamiltonian dynamics with a non-Newtonian momentum. In contrast to MCMC approaches like Hamiltonian Monte Carlo, no stochastic step is required. Instead, the proposed deterministic dynamics in an extended state space exactly sample the target distribution, specified by an energy function, under an assumption of ergodicity. Alternatively, the dynamics can be interpreted as a normalizing flow that samples a specified energy model without training. The proposed Energy Sampling Hamiltonian (ESH) dynamics have a simple form that can be solved with existing ODE solvers, but we derive a specialized solver that exhibits much better performance. ESH dynamics converge faster than their MCMC competitors enabling faster, more stable training of neural network energy models.
****
Dynamic Normalization and Relay for Video Action Recognition

Dongqi Cai, Anbang Yao, Yurong Chen
Convolutional Neural Networks (CNNs) have been the dominant model for video acti
on recognition. Due to the huge memory and compute demand, popular action recogn
ition networks need to be trained with small batch sizes, which makes learning d
iscriminative spatial-temporal representations for videos become a challenging p
roblem. In this paper, we present Dynamic Normalization and Relay (DNR), an impr
oved normalization design, to augment the spatial-temporal representation learni
ng of any deep action recognition model, adapting to small batch size training s
ettings. We observe that state-of-the-art action recognition networks usually ap
ply the same normalization parameters to all video data, and ignore the dependen
cies of the estimated normalization parameters between neighboring frames (at th
e same layer) and between neighboring layers (with all frames of a video clip).
Inspired by this, DNR introduces two dynamic normalization relay modules to expl
ore the potentials of cross-temporal and cross-layer feature distribution depend
encies for estimating accurate layer-wise normalization parameters. These two DN
R modules are instantiated as a light-weight recurrent structure conditioned on
the current input features, and the normalization parameters estimated from the
neighboring frames based features at the same layer or from the whole video clip
 based features at the preceding layers. We first plug DNR into prevailing 2D CN
N backbones and test its performance on public action recognition datasets inclu
ding Kinetics and Something-Something. Experimental results show that DNR brings
 large performance improvements to the baselines, achieving over 4.4% absolute m
argins in top-1 accuracy without training bells and whistles. More experiments o
n 3D backbones and several latest 2D spatial-temporal networks further validate
its effectiveness. Code will be available at https://github.com/caidonkey/dnr.
****
Robust Visual Reasoning via Language Guided Neural Module Networks
Arjun Akula, Varun Jampani, Soravit Changpinyo, Song-Chun Zhu
Neural module networks (NMN) are a popular approach for solving multi-modal task
s such as visual question answering (VQA) and visual referring expression recogn
ition (REF). A key limitation in prior implementations of NMN is that the neural
 modules do not effectively capture the association between the visual input and
 the relevant neighbourhood context of the textual input. This limits their gene
ralizability. For instance, NMN fail to understand new concepts such as “yellow
sphere to the left" even when it is a combination of known concepts from train d
ata: “blue sphere", “yellow cube", and “metallic cube to the left". In this pape
r, we address this limitation by introducing a language-guided adaptive convolut
ion layer (LG-Conv) into NMN, in which the filter weights of convolutions are ex
plicitly multiplied with a spatially varying language-guided kernel. Our model a
llows the neural module to adaptively co-attend over potential objects of intere
st from the visual and textual inputs. Extensive experiments on VQA and REF task
s demonstrate the effectiveness of our approach. Additionally, we propose a new
challenging out-of-distribution test split for REF task, which we call C3-Ref+,
for explicitly evaluating the NMN’s ability to generalize well to adversarial pe
rturbations and unseen combinations of known concepts. Experiments on C3-Ref+ fu
rther demonstrate the generalization capabilities of our approach.
****
True Few-Shot Learning with Language Models
Ethan Perez, Douwe Kiela, Kyunghyun Cho
Pretrained language models (LMs) perform well on many tasks even when learning f
rom a few examples, but prior work uses many held-out examples to tune various a
spects of learning, such as hyperparameters, training objectives, and natural la
nguage templates ("prompts"). Here, we evaluate the few-shot ability of LMs when
 such held-out examples are unavailable, a setting we call true few-shot learnin
g. We test two model selection criteria, cross-validation and minimum descriptio
n length, for choosing LM prompts and hyperparameters in the true few-shot setti
ng. On average, both marginally outperform random selection and greatly underper
form selection based on held-out examples. Moreover, selection criteria often pr
efer models that perform significantly worse than randomly-selected ones. We fin
d similar results even when taking into account our uncertainty in a model's tru

e performance during selection, as well as when varying the amount of computation and number of examples used for selection. Overall, our findings suggest that prior work significantly overestimated the true few-shot ability of LMs given the difficulty of few-shot model selection.
****

Selective Sampling for Online Best-arm Identification
Romain Camilleri, Zhihan Xiong, Maryam Fazel, Lalit Jain, Kevin G. Jamieson
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

Multi-task Learning of Order-Consistent Causal Graphs
Xinshi Chen, Haoran Sun, Caleb Ellington, Eric Xing, Le Song
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

Learning to Iteratively Solve Routing Problems with Dual-Aspect Collaborative Transformer
Yining Ma, Jingwen Li, Zhiguang Cao, Wen Song, Le Zhang, Zhenghua Chen, Jing Tang
Recently, Transformer has become a prevailing deep architecture for solving vehicle routing problems (VRPs). However, it is less effective in learning improvement models for VRP because its positional encoding (PE) method is not suitable in representing VRP solutions. This paper presents a novel Dual-Aspect Collaborative Transformer (DACT) to learn embeddings for the node and positional features separately, instead of fusing them together as done in existing ones, so as to avoid potential noises and incompatible correlations. Moreover, the positional features are embedded through a novel cyclic positional encoding (CPE) method to allow Transformer to effectively capture the circularity and symmetry of VRP solutions (i.e., cyclic sequences). We train DACT using Proximal Policy Optimization and design a curriculum learning strategy for better sample efficiency. We apply DACT to solve the traveling salesman problem (TSP) and capacitated vehicle routing problem (CVRP). Results show that our DACT outperforms existing Transformer based improvement models, and exhibits much better generalization performance across different problem sizes on synthetic and benchmark instances, respectively.
****

Learning interaction rules from multi-animal trajectories via augmented behavioral models
Keisuke Fujii, Naoya Takeishi, Kazushi Tsutsui, Emyo Fujioka, Nozomi Nishiumi, Ryoya Tanaka, Mika Fukushiro, Kaoru Ide, Hiroyoshi Kohno, Ken Yoda, Susumu Takahashi, Shizuko Hiryu, Yoshinobu Kawahara
Extracting the interaction rules of biological agents from movement sequences pose challenges in various domains. Granger causality is a practical framework for analyzing the interactions from observed time-series data; however, this framework ignores the structures and assumptions of the generative process in animal behaviors, which may lead to interpretational problems and sometimes erroneous assessments of causality. In this paper, we propose a new framework for learning Granger causality from multi-animal trajectories via augmented theory-based behavioral models with interpretable data-driven models. We adopt an approach for augmenting incomplete multi-agent behavioral models described by time-varying dynamical systems with neural networks. For efficient and interpretable learning, our model leverages theory-based architectures separating navigation and motion processes, and the theory-guided regularization for reliable behavioral modeling. This can provide interpretable signs of Granger-causal effects over time, i.e., when specific others cause the approach or separation. In experiments using synthetic datasets, our method achieved better performance than various baselines. We then analyzed multi-animal datasets of mice, flies, birds, and bats, which veri

fied our method and obtained novel biological insights.
****
Differentiable Synthesis of Program Architectures
Guofeng Cui, He Zhu
Differentiable programs have recently attracted much interest due to their interpretability, compositionality, and their efficiency to leverage differentiable training. However, synthesizing differentiable programs requires optimizing over a combinatorial, rapidly exploded space of program architectures. Despite the development of effective pruning heuristics, previous works essentially enumerate the discrete search space of program architectures, which is inefficient. We propose to encode program architecture search as learning the probability distribution over all possible program derivations induced by a context-free grammar. This allows the search algorithm to efficiently prune away unlikely program derivations to synthesize optimal program architectures. To this end, an efficient gradient-descent based method is developed to conduct program architecture search in a continuous relaxation of the discrete space of grammar rules. Experiment results on four sequence classification tasks demonstrate that our program synthesizer excels in discovering program architectures that lead to differentiable programs with higher F1 scores, while being more efficient than state-of-the-art program synthesis methods.
****
Make Sure You're Unsure: A Framework for Verifying Probabilistic Specifications
Leonard Berrada, Sumanth Dathathri, Krishnamurthy Dvijotham, Robert Stanforth, Rudy R. Bunel, Jonathan Uesato, Sven Gowal, M. Pawan 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.
****
Oracle-Efficient Regret Minimization in Factored MDPs with Unknown Structure
Aviv Rosenberg, Yishay Mansour
We study regret minimization in non-episodic factored Markov decision processes (FMDPs), where all existing algorithms make the strong assumption that the factored structure of the FMDP is known to the learner in advance. In this paper, we provide the first algorithm that learns the structure of the FMDP while minimizing the regret. Our algorithm is based on the optimism in face of uncertainty principle, combined with a simple statistical method for structure learning, and can be implemented efficiently given oracle-access to an FMDP planner. Moreover, we give a variant of our algorithm that remains efficient even when the oracle is limited to non-factored actions, which is the case with almost all existing approximate planners. Finally, we leverage our techniques to prove a novel lower bound for the known structure case, closing the gap to the regret bound of Chen et al. [2021].
****
Linear-Time Probabilistic Solution of Boundary Value Problems
Nicholas Krämer, Philipp Hennig
We propose a fast algorithm for the probabilistic solution of boundary value problems (BVPs), which are ordinary differential equations subject to boundary conditions. In contrast to previous work, we introduce a Gauss-Markov prior and tailor it specifically to BVPs, which allows computing a posterior distribution over the solution in linear time, at a quality and cost comparable to that of well-established, non-probabilistic methods. Our model further delivers uncertainty quantification, mesh refinement, and hyperparameter adaptation. We demonstrate how these practical considerations positively impact the efficiency of the scheme. Altogether, this results in a practically usable probabilistic BVP solver that is (in contrast to non-probabilistic algorithms) natively compatible with other parts of the statistical modelling tool-chain.
****
Lifelong Domain Adaptation via Consolidated Internal Distribution
Mohammad Rostami

We develop an algorithm to address unsupervised domain adaptation (UDA) in conti nual learning (CL) settings. The goal is to update a model continually to learn distributional shifts across sequentially arriving tasks with unlabeled data wh ile retaining the knowledge about the past learned tasks. Existing UDA algorit hms address the challenge of domain shift, but they require simultaneous access to the datasets of the source and the target domains. On the other hand, existin g works on CL can handle tasks with labeled data. Our solution is based on cons olidating the learned internal distribution for improved model generalization on new domains and benefitting from experience replay to overcome catastrophic for getting.
****
Counterbalancing Learning and Strategic Incentives in Allocation Markets
Jamie Kang, Faidra Monachou, Moran Koren, Itai Ashlagi
Motivated by the high discard rate of donated organs in the United States, we st udy an allocation problem in the presence of learning and strategic incentives. We consider a setting where a benevolent social planner decides whether and how to allocate a single indivisible object to a queue of strategic agents. The obj ect has a common true quality, good or bad, which is ex-ante unknown to everyon e. Each agent holds an informative, yet noisy, private signal about the quality. To make a correct allocation decision the planner attempts to learn the object quality by truthfully eliciting agents' signals. Under the commonly applied sequ ential offering mechanism, we show that learning is hampered by the presence of strategic incentives as herding may emerge. This can result in incorrect allocat ion and welfare loss. To overcome these issues, we propose a novel class of ince ntive-compatible mechanisms. Our mechanism involves a batch-by-batch, dynamic vo ting process using a majority rule. We prove that the proposed voting mechanisms improve the probability of correct allocation whenever agents are sufficiently well informed. Particularly, we show that such an improvement can be achieved vi a a simple greedy algorithm. We quantify the improvement using simulations.
****
Controlling Neural Networks with Rule Representations
Sungyong Seo, Sercan Arik, Jinsung Yoon, Xiang Zhang, Kihyuk Sohn, Tomas Pfister
We propose a novel training method that integrates rules into deep learning, in a way the strengths of the rules are controllable at inference. Deep Neural Netw orks with Controllable Rule Representations (DeepCTRL) incorporates a rule encod er into the model coupled with a rule-based objective, enabling a shared represe ntation for decision making. DeepCTRL is agnostic to data type and model archite cture. It can be applied to any kind of rule defined for inputs and outputs. The key aspect of DeepCTRL is that it does not require retraining to adapt the rule strength -- at inference, the user can adjust it based on the desired operation point on accuracy vs. rule verification ratio. In real-world domains where inco rporating rules is critical -- such as Physics, Retail and Healthcare -- we show the effectiveness of DeepCTRL in teaching rules for deep learning. DeepCTRL imp roves the trust and reliability of the trained models by significantly increasin g their rule verification ratio, while also providing accuracy gains at downstre am tasks. Additionally, DeepCTRL enables novel use cases such as hypothesis test ing of the rules on data samples, and unsupervised adaptation based on shared ru les between datasets.
****
Making the most of your day: online learning for optimal allocation of time
Etienne Boursier, Tristan Garrec, Vianney Perchet, Marco Scarsini
We study online learning for optimal allocation when the resource to be allocate d is time. An agent receives task proposals sequentially according to a Poisson process and can either accept or reject a proposed task. If she accepts the prop osal, she is busy for the duration of the task and obtains a reward that depends on the task duration. If she rejects it, she remains on hold until a new task p roposal arrives. We study the regret incurred by the agent first when she knows her reward function but does not know the distribution of the task duration, and then when she does not know her reward function, either. Faster rates are final ly obtained by adding structural assumptions on the distribution of rides or on

the reward function. This natural setting bears similarities with contextual (one-armed) bandits, but with the crucial difference that the normalized reward associated to a context depends on the whole distribution of contexts.
****

Federated Reconstruction: Partially Local Federated Learning
Karan Singhal, Hakim Sidahmed, Zachary Garrett, Shanshan Wu, John Rush, Sushant Prakash
Personalization methods in federated learning aim to balance the benefits of federated and local training for data availability, communication cost, and robustness to client heterogeneity. Approaches that require clients to communicate all model parameters can be undesirable due to privacy and communication constraints. Other approaches require always-available or stateful clients, impractical in large-scale cross-device settings. We introduce Federated Reconstruction, the first model-agnostic framework for partially local federated learning suitable for training and inference at scale. We motivate the framework via a connection to model-agnostic meta learning, empirically demonstrate its performance over existing approaches for collaborative filtering and next word prediction, and release an open-source library for evaluating approaches in this setting. We also describe the successful deployment of this approach at scale for federated collaborative filtering in a mobile keyboard application.
****

Optimal prediction of Markov chains with and without spectral gap
Yanjun Han, Soham Jana, Yihong 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.
****

Subquadratic Overparameterization for Shallow Neural Networks
ChaeHwan Song, Ali Ramezani-Kebrya, Thomas Pethick, Armin Eftekhari, Volkan Cevher
Overparameterization refers to the important phenomenon where the width of a neural network is chosen such that learning algorithms can provably attain zero loss in nonconvex training. The existing theory establishes such global convergence using various initialization strategies, training modifications, and width scalings. In particular, the state-of-the-art results require the width to scale quadratically with the number of training data under standard initialization strategies used in practice for best generalization performance. In contrast, the most recent results obtain linear scaling either with requiring initializations that lead to the "lazy-training", or training only a single layer. In this work, we provide an analytical framework that allows us to adopt standard initialization strategies, possibly avoid lazy training, and train all layers simultaneously in basic shallow neural networks while attaining a desirable subquadratic scaling on the network width. We achieve the desiderata via Polyak-Lojasiewicz condition, smoothness, and standard assumptions on data, and use tools from random matrix theory.
****

Continuous Doubly Constrained Batch Reinforcement Learning
Rasool Fakoor, Jonas W. Mueller, Kavosh Asadi, Pratik Chaudhari, Alexander J. Smola
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

Bridging Explicit and Implicit Deep Generative Models via Neural Stein Estimators
Qitian Wu, Rui Gao, Hongyuan Zha
There are two types of deep generative models: explicit and implicit. The former defines an explicit density form that allows likelihood inference; while the la

tter targets a flexible transformation from random noise to generated samples. While the two classes of generative models have shown great power in many applications, both of them, when used alone, suffer from respective limitations and drawbacks. To take full advantages of both models and enable mutual compensation, we propose a novel joint training framework that bridges an explicit (unnormalized) density estimator and an implicit sample generator via Stein discrepancy. We show that our method 1) induces novel mutual regularization via kernel Sobolev norm penalization and Moreau-Yosida regularization, and 2) stabilizes the training dynamics. Empirically, we demonstrate that proposed method can facilitate the density estimator to more accurately identify data modes and guide the generator to output higher-quality samples, comparing with training a single counterpart. The new approach also shows promising results when the training samples are contaminated or limited.
****
Score-based Generative Modeling in Latent Space
Arash Vahdat, Karsten Kreis, Jan Kautz
Score-based generative models (SGMs) have recently demonstrated impressive results in terms of both sample quality and distribution coverage. However, they are usually applied directly in data space and often require thousands of network evaluations for sampling. Here, we propose the Latent Score-based Generative Model (LSGM), a novel approach that trains SGMs in a latent space, relying on the variational autoencoder framework. Moving from data to latent space allows us to train more expressive generative models, apply SGMs to non-continuous data, and learn smoother SGMs in a smaller space, resulting in fewer network evaluations and faster sampling. To enable training LSGMs end-to-end in a scalable and stable manner, we (i) introduce a new score-matching objective suitable to the LSGM setting, (ii) propose a novel parameterization of the score function that allows SGM to focus on the mismatch of the target distribution with respect to a simple Normal one, and (iii) analytically derive multiple techniques for variance reduction of the training objective. LSGM obtains a state-of-the-art FID score of 2.10 on CIFAR-10, outperforming all existing generative results on this dataset. On CelebA-HQ-256, LSGM is on a par with previous SGMs in sample quality while outperforming them in sampling time by two orders of magnitude. In modeling binary images, LSGM achieves state-of-the-art likelihood on the binarized OMNIGLOT dataset.
****
Deep Conditional Gaussian Mixture Model for Constrained Clustering
Laura Manduchi, Kieran Chin-Cheong, Holger Michel, Sven Wellmann, Julia Vogt
Constrained clustering has gained significant attention in the field of machine learning as it can leverage prior information on a growing amount of only partially labeled data. Following recent advances in deep generative models, we propose a novel framework for constrained clustering that is intuitive, interpretable, and can be trained efficiently in the framework of stochastic gradient variational inference. By explicitly integrating domain knowledge in the form of probabilistic relations, our proposed model (DC-GMM) uncovers the underlying distribution of data conditioned on prior clustering preferences, expressed as \textit{pairwise constraints}. These constraints guide the clustering process towards a desirable partition of the data by indicating which samples should or should not belong to the same cluster. We provide extensive experiments to demonstrate that DC-GMM shows superior clustering performances and robustness compared to state-of-the-art deep constrained clustering methods on a wide range of data sets. We further demonstrate the usefulness of our approach on two challenging real-world applications.
****
Bootstrap Your Object Detector via Mixed Training
Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Stephen Lin, Han Hu, Xiang Bai
We introduce MixTraining, a new training paradigm for object detection that can improve the performance of existing detectors for free. MixTraining enhances data augmentation by utilizing augmentations of different strengths while excluding

the strong augmentations of certain training samples that may be detrimental to training. In addition, it addresses localization noise and missing labels in human annotations by incorporating pseudo boxes that can compensate for these errors. Both of these MixTraining capabilities are made possible through bootstrapping on the detector, which can be used to predict the difficulty of training on a strong augmentation, as well as to generate reliable pseudo boxes thanks to the robustness of neural networks to labeling error. MixTraining is found to bring consistent improvements across various detectors on the COCO dataset. In particular, the performance of Faster R-CNN~\cite{ren2015faster} with a ResNet-50~\cite{he2016deep} backbone is improved from 41.7 mAP to 44.0 mAP, and the accuracy of Cascade-RCNN~\cite{cai2018cascade} with a Swin-Small~\cite{liu2021swin} backbone is raised from 50.9 mAP to 52.8 mAP.
****
Tensor decompositions of higher-order correlations by nonlinear Hebbian plasticity
Gabriel Ocker, Michael Buice
Biological synaptic plasticity exhibits nonlinearities that are not accounted for by classic Hebbian learning rules. Here, we introduce a simple family of generalized nonlinear Hebbian learning rules. We study the computations implemented by their dynamics in the simple setting of a neuron receiving feedforward inputs. These nonlinear Hebbian rules allow a neuron to learn tensor decompositions of its higher- order input correlations. The particular input correlation decomposed and the form of the decomposition depend on the location of nonlinearities in the plasticity rule. For simple, biologically motivated parameters, the neuron learns eigenvectors of higher-order input correlation tensors. We prove that tensor eigenvectors are attractors and determine their basins of attraction. We calculate the volume of those basins, showing that the dominant eigenvector has the largest basin of attraction. We then study arbitrary learning rules and find that any learning rule that admits a finite Taylor expansion into the neural input and output also has stable equilibria at generalized eigenvectors of higher-order input correlation tensors. Nonlinearities in synaptic plasticity thus allow a neuron to encode higher-order input correlations in a simple fashion.
****
Online Adaptation to Label Distribution Shift
Ruihan Wu, Chuan Guo, Yi Su, Kilian Q. Weinberger
Machine learning models often encounter distribution shifts when deployed in the real world. In this paper, we focus on adaptation to label distribution shift in the online setting, where the test-time label distribution is continually changing and the model must dynamically adapt to it without observing the true label. This setting is common in many real world scenarios such as medical diagnosis, where disease prevalences can vary substantially at different times of the year. Leveraging a novel analysis, we show that the lack of true label does not hinder estimation of the expected test loss, which enables the reduction of online label shift adaptation to conventional online learning. Informed by this observation, we propose adaptation algorithms inspired by classical online learning techniques such as Follow The Leader (FTL) and Online Gradient Descent (OGD) and derive their regret bounds. We empirically verify our findings under both simulated and real world label distribution shifts and show that OGD is particularly effective and robust to a variety of challenging label shift scenarios.
****
One Explanation is Not Enough: Structured Attention Graphs for Image Classification
Vivswan Shitole, Fuxin Li, Minsuk Kahng, Prasad Tadepalli, Alan Fern
Attention maps are popular tools for explaining the decisions of convolutional neural networks (CNNs) for image classification. Typically, for each image of interest, a single attention map is produced, which assigns weights to pixels based on their importance to the classification. We argue that a single attention map provides an incomplete understanding since there are often many other maps that explain a classification equally well. In this paper, we propose to utilize a beam search algorithm to systematically search for multiple explanations for each

image. Results show that there are indeed multiple relatively localized explanations for many images. However, naively showing multiple explanations to users can be overwhelming and does not reveal their common and distinct structures. We introduce structured attention graphs (SAGs), which compactly represent sets of attention maps for an image by visualizing how different combinations of image regions impact the confidence of a classifier. An approach to computing a compact and representative SAG for visualization is proposed via diverse sampling. We conduct a user study comparing the use of SAGs to traditional attention maps for answering comparative counterfactual questions about image classifications. Our results show that the users are significantly more accurate when presented with SAGs compared to standard attention map baselines.
****
Integrating Expert ODEs into Neural ODEs: Pharmacology and Disease Progression
Zhaozhi Qian, William Zame, Lucas Fleuren, Paul Elbers, Mihaela van der Schaar
Modeling a system's temporal behaviour in reaction to external stimuli is a fundamental problem in many areas. Pure Machine Learning (ML) approaches often fail in the small sample regime and cannot provide actionable insights beyond predictions. A promising modification has been to incorporate expert domain knowledge into ML models. The application we consider is predicting the patient health status and disease progression over time, where a wealth of domain knowledge is available from pharmacology. Pharmacological models describe the dynamics of carefully-chosen medically meaningful variables in terms of systems of Ordinary Differential Equations (ODEs). However, these models only describe a limited collection of variables, and these variables are often not observable in clinical environments. To close this gap, we propose the latent hybridisation model (LHM) that integrates a system of expert-designed ODEs with machine-learned Neural ODEs to fully describe the dynamics of the system and to link the expert and latent variables to observable quantities. We evaluated LHM on synthetic data as well as real-world intensive care data of COVID-19 patients. LHM consistently outperforms previous works, especially when few training samples are available such as at the beginning of the pandemic.
****
Shifted Chunk Transformer for Spatio-Temporal Representational Learning
Xuefan Zha, Wentao Zhu, Lv Xun, Sen Yang, Ji Liu
Spatio-temporal representational learning has been widely adopted in various fields such as action recognition, video object segmentation, and action anticipation.Previous spatio-temporal representational learning approaches primarily employ ConvNets or sequential models, e.g., LSTM, to learn the intra-frame and inter-frame features. Recently, Transformer models have successfully dominated the study of natural language processing (NLP), image classification, etc. However, the pure-Transformer based spatio-temporal learning can be prohibitively costly on memory and computation to extract fine-grained features from a tiny patch. To tackle the training difficulty and enhance the spatio-temporal learning, we construct a shifted chunk Transformer with pure self-attention blocks. Leveraging the recent efficient Transformer design in NLP, this shifted chunk Transformer can learn hierarchical spatio-temporal features from a local tiny patch to a global videoclip. Our shifted self-attention can also effectively model complicated inter-frame variances. Furthermore, we build a clip encoder based on Transformer to model long-term temporal dependencies. We conduct thorough ablation studies to validate each component and hyper-parameters in our shifted chunk Transformer, and it outperforms previous state-of-the-art approaches on Kinetics-400, Kinetics-600,UCF101, and HMDB51.
****
Faster proximal algorithms for matrix optimization using Jacobi-based eigenvalue methods
Hamza Fawzi, Harry Goulbourne
We consider proximal splitting algorithms for convex optimization problems over matrices. A significant computational bottleneck in many of these algorithms is the need to compute a full eigenvalue or singular value decomposition at each iteration for the evaluation of a proximal operator.In this paper we propose to us

e an old and surprisingly simple method due to Jacobi to compute these eigenvalu
e and singular value decompositions, and we demonstrate that it can lead to subs
tantial gains in terms of computation time compared to standard approaches. We r
ely on three essential properties of this method: (a) its ability to exploit an
approximate decomposition as an initial point, which in the case of iterative op
timization algorithms can be obtained from the previous iterate; (b) its paralle
l nature which makes it a great fit for hardware accelerators such as GPUs, now
common in machine learning, and (c) its simple termination criterion which allow
s us to trade-off accuracy with computation time. We demonstrate the efficacy of
 this approach on a variety of algorithms and problems, and show that, on a GPU,
 we can obtain 5 to 10x speed-ups in the evaluation of proximal operators compar
ed to standard CPU or GPU linear algebra routines. Our findings are supported by
 new theoretical results providing guarantees on the approximation quality of pr
oximal operators obtained using approximate eigenvalue or singular value decompo
sitions.
****
Decrypting Cryptic Crosswords: Semantically Complex Wordplay Puzzles as a Target
 for NLP
Josh Rozner, Christopher Potts, Kyle Mahowald
Cryptic crosswords, the dominant crossword variety in the UK, are a promising ta
rget for advancing NLP systems that seek to process semantically complex, highly
 compositional language. Cryptic clues read like fluent natural language but are
 adversarially composed of two parts: a definition and a wordplay cipher requiri
ng character-level manipulations. Expert humans use creative intelligence to sol
ve cryptics, flexibly combining linguistic, world, and domain knowledge. In this
 paper, we make two main contributions. First, we present a dataset of cryptic c
lues as a challenging new benchmark for NLP systems that seek to process composi
tional language in more creative, human-like ways. After showing that three non-
neural approaches and T5, a state-of-the-art neural language model, do not achie
ve good performance, we make our second main contribution: a novel curriculum ap
proach, in which the model is first fine-tuned on related tasks such as unscramb
ling words. We also introduce a challenging data split, examine the meta-linguis
tic capabilities of subword-tokenized models, and investigate model systematicit
y by perturbing the wordplay part of clues, showing that T5 exhibits behavior pa
rtially consistent with human solving strategies. Although our curricular approa
ch considerably improves on the T5 baseline, our best-performing model still fai
ls to generalize to the extent that humans can. Thus, cryptic crosswords remain
an unsolved challenge for NLP systems and a potential source of future innovatio
n.
****
An Improved Analysis of Gradient Tracking for Decentralized Machine Learning
Anastasiia Koloskova, Tao Lin, Sebastian U. Stich
Requests for name changes in the electronic proceedings will be accepted with 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.
****
Entropic Desired Dynamics for Intrinsic Control
Steven Hansen, Guillaume Desjardins, Kate Baumli, David Warde-Farley, Nicolas He
ess, Simon Osindero, Volodymyr Mnih
An agent might be said, informally, to have mastery of its environment when it h
as maximised the effective number of states it can reliably reach. In practice,
this often means maximizing the number of latent codes that can be discriminated
 from future states under some short time horizon (e.g. \cite{eysenbach2018diver
sity}). By situating these latent codes in a globally consistent coordinate syst
em, we show that agents can reliably reach more states in the long term while st
ill optimizing a local objective. A simple instantiation of this idea, \textbf{E
}ntropic \textbf{D}esired \textbf{D}ynamics for \textbf{I}ntrinsic \textbf{C}on\
textbf{T}rol (EDDICT), assumes fixed additive latent dynamics, which results in
tractable learning and an interpretable latent space. Compared to prior methods,

EDDICT's globally consistent codes allow it to be far more exploratory, as demonstrated by improved state coverage and increased unsupervised performance on hard exploration games such as Montezuma's Revenge.
****

Exploring Cross-Video and Cross-Modality Signals for Weakly-Supervised Audio-Visual Video Parsing
Yan-Bo Lin, Hung-Yu Tseng, Hsin-Ying Lee, Yen-Yu Lin, Ming-Hsuan Yang
The audio-visual video parsing task aims to temporally parse a video into audio or visual event categories. However, it is labor intensive to temporally annotate audio and visual events and thus hampers the learning of a parsing model. To this end, we propose to explore additional cross-video and cross-modality supervisory signals to facilitate weakly-supervised audio-visual video parsing. The proposed method exploits both the common and diverse event semantics across videos to identify audio or visual events. In addition, our method explores event co-occurrence across audio, visual, and audio-visual streams. We leverage the explored cross-modality co-occurrence to localize segments of target events while excluding irrelevant ones. The discovered supervisory signals across different videos and modalities can greatly facilitate the training with only video-level annotations. Quantitative and qualitative results demonstrate that the proposed method performs favorably against existing methods on weakly-supervised audio-visual video parsing.
****

Littlestone Classes are Privately Online Learnable
Noah Golowich, Roi Livni
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 Parameterization of Sparse Variational Gaussian Processes
Vincent ADAM, Paul Chang, Mohammad Emtiyaz E. Khan, Arno Solin
Sparse variational Gaussian process (SVGP) methods are a common choice for non-conjugate Gaussian process inference because of their computational benefits. In this paper, we improve their computational efficiency by using a dual parameterization where each data example is assigned dual parameters, similarly to site parameters used in expectation propagation. Our dual parameterization speeds-up inference using natural gradient descent, and provides a tighter evidence lower bound for hyperparameter learning. The approach has the same memory cost as the current SVGP methods, but it is faster and more accurate.
****

Learning to dehaze with polarization
Chu Zhou, Minggui Teng, Yufei Han, Chao Xu, Boxin Shi
Haze, a common kind of bad weather caused by atmospheric scattering, decreases the visibility of scenes and degenerates the performance of computer vision algorithms. Single-image dehazing methods have shown their effectiveness in a large variety of scenes, however, they are based on handcrafted priors or learned features, which do not generalize well to real-world images. Polarization information can be used to relieve its ill-posedness, however, real-world images are still challenging since existing polarization-based methods usually assume that the transmitted light is not significantly polarized, and they require specific clues to estimate necessary physical parameters. In this paper, we propose a generalized physical formation model of hazy images and a robust polarization-based dehazing pipeline without the above assumption or requirement, along with a neural network tailored to the pipeline. Experimental results show that our approach achieves state-of-the-art performance on both synthetic data and real-world hazy images.
****

Conservative Data Sharing for Multi-Task Offline Reinforcement Learning
Tianhe Yu, Aviral Kumar, Yevgen Chebotar, Karol Hausman, Sergey Levine, Chelsea Finn

Offline reinforcement learning (RL) algorithms have shown promising results in domains where abundant pre-collected data is available. However, prior methods focus on solving individual problems from scratch with an offline dataset without considering how an offline RL agent can acquire multiple skills. We argue that a natural use case of offline RL is in settings where we can pool large amounts of data collected in various scenarios for solving different tasks, and utilize all of this data to learn behaviors for all the tasks more effectively rather than training each one in isolation. However, sharing data across all tasks in multi-task offline RL performs surprisingly poorly in practice. Thorough empirical analysis, we find that sharing data can actually exacerbate the distributional shift between the learned policy and the dataset, which in turn can lead to divergence of the learned policy and poor performance. To address this challenge, we develop a simple technique for data- sharing in multi-task offline RL that routes data based on the improvement over the task-specific data. We call this approach conservative data sharing (CDS), and it can be applied with multiple single-task offline RL methods. On a range of challenging multi-task locomotion, navigation, and vision-based robotic manipulation problems, CDS achieves the best or comparable performance compared to prior offline multi- task RL methods and previous data sharing approaches.
****

Universal Rate-Distortion-Perception Representations for Lossy Compression
George Zhang, Jingjing Qian, Jun Chen, Ashish Khisti
In the context of lossy compression, Blau \& Michaeli (2019) adopt a mathematical notion of perceptual quality and define the information rate-distortion-perception function, generalizing the classical rate-distortion tradeoff. We consider the notion of universal representations in which one may fix an encoder and vary the decoder to achieve any point within a collection of distortion and perception constraints. We prove that the corresponding information-theoretic universal rate-distortion-perception function is operationally achievable in an approximate sense. Under MSE distortion, we show that the entire distortion-perception tradeoff of a Gaussian source can be achieved by a single encoder of the same rate asymptotically. We then characterize the achievable distortion-perception region for a fixed representation in the case of arbitrary distributions, and identify conditions under which the aforementioned results continue to hold approximately. This motivates the study of practical constructions that are approximately universal across the RDP tradeoff, thereby alleviating the need to design a new encoder for each objective. We provide experimental results on MNIST and SVHN suggesting that on image compression tasks, the operational tradeoffs achieved by machine learning models with a fixed encoder suffer only a small penalty when compared to their variable encoder counterparts.
****

What’s a good imputation to predict with missing values?
Marine Le Morvan, Julie Josse, Erwan Scornet, Gael Varoquaux
How to learn a good predictor on data with missing values? Most efforts focus on first imputing as well as possible and second learning on the completed data to predict the outcome. Yet, this widespread practice has no theoretical grounding. Here we show that for almost all imputation functions, an impute-then-regress procedure with a powerful learner is Bayes optimal. This result holds for all missing-values mechanisms, in contrast with the classic statistical results that require missing-at-random settings to use imputation in probabilistic modeling. Moreover, it implies that perfect conditional imputation is not needed for good prediction asymptotically. In fact, we show that on perfectly imputed data the best regression function will generally be discontinuous, which makes it hard to learn. Crafting instead the imputation so as to leave the regression function unchanged simply shifts the problem to learning discontinuous imputations. Rather, we suggest that it is easier to learn imputation and regression jointly. We propose such a procedure, adapting NeuMiss, a neural network capturing the conditional links across observed and unobserved variables whatever the missing-value pattern. Our experiments confirm that joint imputation and regression through NeuMiss is better than various two step procedures in a finite-sample regime.

****
Replacing Rewards with Examples: Example-Based Policy Search via Recursive Classification
Ben Eysenbach, Sergey Levine, Russ R. Salakhutdinov
Reinforcement learning (RL) algorithms assume that users specify tasks by manually writing down a reward function. However, this process can be laborious and demands considerable technical expertise. Can we devise RL algorithms that instead enable users to specify tasks simply by providing examples of successful outcomes? In this paper, we derive a control algorithm that maximizes the future probability of these successful outcome examples. Prior work has approached similar problems with a two-stage process, first learning a reward function and then optimizing this reward function using another reinforcement learning algorithm. In contrast, our method directly learns a value function from transitions and successful outcomes, without learning this intermediate reward function. Our method therefore requires fewer hyperparameters to tune and lines of code to debug. We show that our method satisfies a new data-driven Bellman equation, where examples take the place of the typical reward function term. Experiments show that our approach outperforms prior methods that learn explicit reward functions.
****
Hierarchical Skills for Efficient Exploration
Jonas Gehring, Gabriel Synnaeve, Andreas Krause, Nicolas Usunier
In reinforcement learning, pre-trained low-level skills have the potential to greatly facilitate exploration. However, prior knowledge of the downstream task is required to strike the right balance between generality (fine-grained control) and specificity (faster learning) in skill design. In previous work on continuous control, the sensitivity of methods to this trade-off has not been addressed explicitly, as locomotion provides a suitable prior for navigation tasks, which have been of foremost interest. In this work, we analyze this trade-off for low-level policy pre-training with a new benchmark suite of diverse, sparse-reward tasks for bipedal robots. We alleviate the need for prior knowledge by proposing a hierarchical skill learning framework that acquires skills of varying complexity in an unsupervised manner. For utilization on downstream tasks, we present a three-layered hierarchical learning algorithm to automatically trade off between general and specific skills as required by the respective task. In our experiments, we show that our approach performs this trade-off effectively and achieves better results than current state-of-the-art methods for end-to-end hierarchical reinforcement learning and unsupervised skill discovery.
****
Evidential Softmax for Sparse Multimodal Distributions in Deep Generative Models
Phil Chen, Mikhal Itkina, Ransalu Senanayake, Mykel J Kochenderfer
Many applications of generative models rely on the marginalization of their high-dimensional output probability distributions. Normalization functions that yield sparse probability distributions can make exact marginalization more computationally tractable. However, sparse normalization functions usually require alternative loss functions for training since the log-likelihood is undefined for sparse probability distributions. Furthermore, many sparse normalization functions often collapse the multimodality of distributions. In this work, we present ev-softmax, a sparse normalization function that preserves the multimodality of probability distributions. We derive its properties, including its gradient in closed-form, and introduce a continuous family of approximations to ev-softmax that have full support and can be trained with probabilistic loss functions such as negative log-likelihood and Kullback-Leibler divergence. We evaluate our method on a variety of generative models, including variational autoencoders and auto-regressive architectures. Our method outperforms existing dense and sparse normalization techniques in distributional accuracy. We demonstrate that ev-softmax successfully reduces the dimensionality of probability distributions while maintaining multimodality.
****
Submodular + Concave
Siddharth Mitra, 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-authors prior to requesting a name change in the electronic proceedings.
****

DeepGEM: Generalized Expectation-Maximization for Blind Inversion
Angela Gao, Jorge Castellanos, Yisong Yue, Zachary Ross, Katherine Bouman
Typically, inversion algorithms assume that a forward model, which relates a source to its resulting measurements, is known and fixed. Using collected indirect measurements and the forward model, the goal becomes to recover the source. When the forward model is unknown, or imperfect, artifacts due to model mismatch occur in the recovery of the source. In this paper, we study the problem of blind inversion: solving an inverse problem with unknown or imperfect knowledge of the forward model parameters. We propose DeepGEM, a variational Expectation-Maximization (EM) framework that can be used to solve for the unknown parameters of the forward model in an unsupervised manner. DeepGEM makes use of a normalizing flow generative network to efficiently capture complex posterior distributions, which leads to more accurate evaluation of the source's posterior distribution used in EM. We showcase the effectiveness of our DeepGEM approach by achieving strong performance on the challenging problem of blind seismic tomography, where we significantly outperform the standard method used in seismology. We also demonstrate the generality of DeepGEM by applying it to a simple case of blind deconvolution.
****

Learning to Generate Visual Questions with Noisy Supervision
Shen Kai, Lingfei Wu, Siliang Tang, Yueting Zhuang, zhen he, Zhuoye Ding, Yun Xiao, Bo Long
The task of visual question generation (VQG) aims to generate human-like neural questions from an image and potentially other side information (e.g., answer type or the answer itself). Existing works often suffer from the severe one image to many questions mapping problem, which generates uninformative and non-referential questions. Recent work has demonstrated that by leveraging double visual and answer hints, a model can faithfully generate much better quality questions. However, visual hints are not available naturally. Despite they proposed a simple rule-based similarity matching method to obtain candidate visual hints, they could be very noisy practically and thus restrict the quality of generated questions. In this paper, we present a novel learning approach for double-hints based VQG, which can be cast as a weakly supervised learning problem with noises. The key rationale is that the salient visual regions of interest can be viewed as a constraint to improve the generation procedure for producing high-quality questions. As a result, given the predicted salient visual regions of interest, we can focus on estimating the probability of being ground-truth questions, which in turn implicitly measures the quality of predicted visual hints. Experimental results on two benchmark datasets show that our proposed method outperforms the state-of-the-art approaches by a large margin on a variety of metrics, including both automatic machine metrics and human evaluation.
****

Pure Exploration in Kernel and Neural Bandits
Yinglun Zhu, Dongruo Zhou, Ruoxi Jiang, Quanquan Gu, Rebecca Willett, Robert Nowak
We study pure exploration in bandits, where the dimension of the feature representation can be much larger than the number of arms. To overcome the curse of dimensionality, we propose to adaptively embed the feature representation of each arm into a lower-dimensional space and carefully deal with the induced model misspecifications. Our approach is conceptually very different from existing works that can either only handle low-dimensional linear bandits or passively deal with model misspecifications. We showcase the application of our approach to two pure exploration settings that were previously under-studied: (1) the reward function belongs to a possibly infinite-dimensional Reproducing Kernel Hilbert Space, and (2) the reward function is nonlinear and can be approximated by neural netwo

rks. Our main results provide sample complexity guarantees that only depend on the effective dimension of the feature spaces in the kernel or neural representations. Extensive experiments conducted on both synthetic and real-world datasets demonstrate the efficacy of our methods.
****
Numerical Composition of Differential Privacy
Sivakanth Gopi, Yin Tat Lee, Lukas Wutschitz
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Coresets for Classification – Simplified and Strengthened
Tung Mai, Cameron Musco, Anup Rao
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Sequential Algorithms for Testing Closeness of Distributions
Aadil Oufkir, Omar Fawzi, Nicolas Flammarion, Aurélien Garivier
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Overlapping Spaces for Compact Graph Representations
Kirill Shevkunov, Liudmila Prokhorenkova
Various non-trivial spaces are becoming popular for embedding structured data such as graphs, texts, or images. Following spherical and hyperbolic spaces, more general product spaces have been proposed. However, searching for the best configuration of a product space is a resource-intensive procedure, which reduces the practical applicability of the idea. We generalize the concept of product space and introduce an overlapping space that does not have the configuration search problem. The main idea is to allow subsets of coordinates to be shared between spaces of different types (Euclidean, hyperbolic, spherical). As a result, we often need fewer coordinates to store the objects. Additionally, we propose an optimization algorithm that automatically learns the optimal configuration. Our experiments confirm that overlapping spaces outperform the competitors in graph embedding tasks with different evaluation metrics. We also perform an empirical analysis in a realistic information retrieval setup, where we compare all spaces by incorporating them into DSSM. In this case, the proposed overlapping space consistently achieves nearly optimal results without any configuration tuning. This allows for reducing training time, which can be essential in large-scale applications.
****
Hyperparameter Tuning is All You Need for LISTA
Xiaohan Chen, Jialin Liu, Zhangyang Wang, Wotao Yin
Learned Iterative Shrinkage-Thresholding Algorithm (LISTA) introduces the concept of unrolling an iterative algorithm and training it like a neural network. It has had great success on sparse recovery. In this paper, we show that adding momentum to intermediate variables in the LISTA network achieves a better convergence rate and, in particular, the network with instance-optimal parameters is superlinearly convergent. Moreover, our new theoretical results lead to a practical approach of automatically and adaptively calculating the parameters of a LISTA network layer based on its previous layers. Perhaps most surprisingly, such an adaptive-parameter procedure reduces the training of LISTA to tuning only three hyperparameters from data: a new record set in the context of the recent advances on trimming down LISTA complexity. We call this new ultra-light weight network HyperLISTA. Compared to state-of-the-art LISTA models, HyperLISTA achieves almost

the same performance on seen data distributions and performs better when tested
on unseen distributions (speci■cally, those with different sparsity levels and
nonzero magnitudes). Code is available: https://github.com/VITA-Group/HyperLISTA
.
****
Foundations of Symbolic Languages for Model Interpretability
Marcelo Arenas, Daniel Báez, Pablo Barceló, Jorge Pérez, Bernardo Subercaseaux
Several queries and scores have recently been proposed to explain individual pre
dictions over ML models. Examples include queries based on “anchors”, which are
parts of an instance that are sufficient to justify its classification, and “fea
ture-perturbation” scores such as SHAP. Given the need for flexible, reliable, a
nd easy-to-apply interpretability methods for ML models, we foresee the need for
developing declarative languages to naturally specify different explainability
queries. We do this in a principled way by rooting such a language in a logic ca
lled FOIL, which allows for expressing many simple but important explainability
queries, and might serve as a core for more expressive interpretability language
s. We study the computational complexity of FOIL queries over two classes of ML
models often deemed to be easily interpretable: decision trees and more general
decision diagrams. Since the number of possible inputs for an ML model is expone
ntial in its dimension, tractability of the FOIL evaluation problem is delicate
but can be achieved by either restricting the structure of the models, or the fr
agment of FOIL being evaluated. We also present a prototype implementation of F
OIL wrapped in a high-level declarative language and perform experiments showing
that such a language can be used in practice.
****
Bridging Offline Reinforcement Learning and Imitation Learning: A Tale of Pessim
ism
Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, Stuart Russell
Requests for name changes in the electronic proceedings will be accepted with 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.
****
Impression learning: Online representation learning with synaptic plasticity
Colin Bredenberg, Benjamin Lyo, Eero Simoncelli, Cristina Savin
Understanding how the brain constructs statistical models of the sensory world r
emains a longstanding challenge for computational neuroscience. Here, we derive
an unsupervised local synaptic plasticity rule that trains neural circuits to in
fer latent structure from sensory stimuli via a novel loss function for approxim
ate online Bayesian inference. The learning algorithm is driven by a local error
signal computed between two factors that jointly contribute to neural activity:
stimulus drive and internal predictions --- the network's 'impression' of the s
timulus. Physiologically, we associate these two components with the basal and a
pical dendrites of pyramidal neurons, respectively. We show that learning can be
implemented online, is capable of capturing temporal dependencies in continuous
input streams, and generalizes to hierarchical architectures. Furthermore, we d
emonstrate both analytically and empirically that the algorithm is more data-eff
icient than a three-factor plasticity alternative, enabling it to learn statisti
cs of high-dimensional, naturalistic inputs. Overall, the model provides a bridg
e from mechanistic accounts of synaptic plasticity to algorithmic descriptions o
f unsupervised probabilistic learning and inference.
****
How Well do Feature Visualizations Support Causal Understanding of CNN Activatio
ns?
Roland S. Zimmermann, Judy Borowski, Robert Geirhos, Matthias Bethge, Thomas Wal
lis, Wieland Brendel
Requests for name changes in the electronic proceedings will be accepted with 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.

****
Fixes That Fail: Self-Defeating Improvements in Machine-Learning Systems
Ruihan Wu, Chuan Guo, Awni Hannun, Laurens van der Maaten
Machine-learning systems such as self-driving cars or virtual assistants are composed of a large number of machine-learning models that recognize image content, transcribe speech, analyze natural language, infer preferences, rank options, etc. Models in these systems are often developed and trained independently, which raises an obvious concern: Can improving a machine-learning model make the overall system worse? We answer this question affirmatively by showing that improving a model can deteriorate the performance of downstream models, even after those downstream models are retrained. Such self-defeating improvements are the result of entanglement between the models in the system. We perform an error decomposition of systems with multiple machine-learning models, which sheds light on the types of errors that can lead to self-defeating improvements. We also present the results of experiments which show that self-defeating improvements emerge in a realistic stereo-based detection system for cars and pedestrians.
****
Coarse-to-fine Animal Pose and Shape Estimation
Chen Li, Gim Hee Lee
Most existing animal pose and shape estimation approaches reconstruct animal meshes with a parametric SMAL model. This is because the low-dimensional pose and shape parameters of the SMAL model makes it easier for deep networks to learn the high-dimensional animal meshes. However, the SMAL model is learned from scans of toy animals with limited pose and shape variations, and thus may not be able to represent highly varying real animals well. This may result in poor fittings of the estimated meshes to the 2D evidences, e.g. 2D keypoints or silhouettes. To mitigate this problem, we propose a coarse-to-fine approach to reconstruct 3D animal mesh from a single image. The coarse estimation stage first estimates the pose, shape and translation parameters of the SMAL model. The estimated meshes are then used as a starting point by a graph convolutional network (GCN) to predict a per-vertex deformation in the refinement stage. This combination of SMAL-based and vertex-based representations benefits from both parametric and non-parametric representations. We design our mesh refinement GCN (MRGCN) as an encoder-decoder structure with hierarchical feature representations to overcome the limited receptive field of traditional GCNs. Moreover, we observe that the global image feature used by existing animal mesh reconstruction works is unable to capture detailed shape information for mesh refinement. We thus introduce a local feature extractor to retrieve a vertex-level feature and use it together with the global feature as the input of the MRGCN. We test our approach on the StanfordExtra dataset and achieve state-of-the-art results. Furthermore, we test the generalization capacity of our approach on the Animal Pose and BADJA datasets. Our code is available at the project website.
****
Meta-Learning Sparse Implicit Neural Representations
Jaeho Lee, Jihoon Tack, Namhoon Lee, Jinwoo Shin
Implicit neural representations are a promising new avenue of representing general signals by learning a continuous function that, parameterized as a neural network, maps the domain of a signal to its codomain; the mapping from spatial coordinates of an image to its pixel values, for example. Being capable of conveying fine details in a high dimensional signal, unboundedly of its domain, implicit neural representations ensure many advantages over conventional discrete representations. However, the current approach is difficult to scale for a large number of signals or a data set, since learning a neural representation---which is parameter heavy by itself---for each signal individually requires a lot of memory and computations. To address this issue, we propose to leverage a meta-learning approach in combination with network compression under a sparsity constraint, such that it renders a well-initialized sparse parameterization that evolves quickly to represent a set of unseen signals in the subsequent training. We empirically demonstrate that meta-learned sparse neural representations achieve a much smaller loss than dense meta-learned models with the same number of parameters, whe

n trained to fit each signal using the same number of optimization steps.
****
Rethinking Space-Time Networks with Improved Memory Coverage for Efficient Video Object Segmentation
Ho Kei Cheng, Yu-Wing Tai, Chi-Keung Tang
This paper presents a simple yet effective approach to modeling space-time correspondences in the context of video object segmentation. Unlike most existing approaches, we establish correspondences directly between frames without re-encoding the mask features for every object, leading to a highly efficient and robust framework. With the correspondences, every node in the current query frame is inferred by aggregating features from the past in an associative fashion. We cast the aggregation process as a voting problem and find that the existing inner-product affinity leads to poor use of memory with a small (fixed) subset of memory nodes dominating the votes, regardless of the query. In light of this phenomenon, we propose using the negative squared Euclidean distance instead to compute the affinities. We validated that every memory node now has a chance to contribute, and experimentally showed that such diversified voting is beneficial to both memory efficiency and inference accuracy. The synergy of correspondence networks and diversified voting works exceedingly well, achieves new state-of-the-art results on both DAVIS and YouTubeVOS datasets while running significantly faster at 20+ FPS for multiple objects without bells and whistles.
****
Sparse Spiking Gradient Descent
Nicolas Perez-Nieves, Dan Goodman
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 Calibration of Deep Neural Networks: Do Not Be Afraid of Overconfidence
Deng-Bao Wang, Lei Feng, Min-Ling Zhang
Capturing accurate uncertainty quantification of the prediction from deep neural networks is important in many real-world decision-making applications. A reliable predictor is expected to be accurate when it is confident about its predictions and indicate high uncertainty when it is likely to be inaccurate. However, modern neural networks have been found to be poorly calibrated, primarily in the direction of overconfidence. In recent years, there is a surge of research on model calibration by leveraging implicit or explicit regularization techniques during training, which obtain well calibration by avoiding overconfident outputs. In our study, we empirically found that despite the predictions obtained from these regularized models are better calibrated, they suffer from not being as calibratable, namely, it is harder to further calibrate their predictions with post-hoc calibration methods like temperature scaling and histogram binning. We conduct a series of empirical studies showing that overconfidence may not hurt final calibration performance if post-hoc calibration is allowed, rather, the penalty of confident outputs will compress the room of potential improvements in post-hoc calibration phase. Our experimental findings point out a new direction to improve calibration of DNNs by considering main training and post-hoc calibration as a unified framework.
****
Towards Efficient and Effective Adversarial Training
Gaurang Sriramanan, Sravanti Addepalli, Arya Baburaj, Venkatesh Babu R
The vulnerability of Deep Neural Networks to adversarial attacks has spurred immense interest towards improving their robustness. However, present state-of-the-art adversarial defenses involve the use of 10-step adversaries during training, which renders them computationally infeasible for application to large-scale datasets. While the recent single-step defenses show promising direction, their robustness is not on par with multi-step training methods. In this work, we bridge this performance gap by introducing a novel Nuclear-Norm regularizer on network

predictions to enforce function smoothing in the vicinity of data samples. While prior works consider each data sample independently, the proposed regularizer uses the joint statistics of adversarial samples across a training minibatch to enhance optimization during both attack generation and training, obtaining state-of-the-art results amongst efficient defenses. We achieve further gains by incorporating exponential averaging of network weights over training iterations. We finally introduce a Hybrid training approach that combines the effectiveness of a two-step variant of the proposed defense with the efficiency of a single-step defense. We demonstrate superior results when compared to multi-step defenses such as TRADES and PGD-AT as well, at a significantly lower computational cost.
****

Intriguing Properties of Contrastive Losses

Ting Chen, Calvin Luo, Lala Li

We study three intriguing properties of contrastive learning. First, we generalize the standard contrastive loss to a broader family of losses, and we find that various instantiations of the generalized loss perform similarly under the presence of a multi-layer non-linear projection head. Second, we study if instance-based contrastive learning (with a global image representation) can learn well on images with multiple objects present. We find that meaningful hierarchical local features can be learned despite the fact that these objectives operate on global instance-level features. Finally, we study the phenomenon of feature suppression among competing features shared across augmented views, such as "color distribution" vs "object class". We construct datasets with explicit and controllable competing features, and show that, for contrastive learning, a few bits of easy-to-learn shared features can suppress, and even fully prevent, the learning of other sets of competing features. In scenarios where there are multiple objects in an image, the dominant object would suppress the learning of smaller objects. Existing contrastive learning methods critically rely on data augmentation to favor certain sets of features over others, and could suffer from learning saturation for scenarios where existing augmentations cannot fully address the feature suppression. This poses open challenges to existing contrastive learning techniques.
****

Detecting Moments and Highlights in Videos via Natural Language Queries

Jie Lei, Tamara L Berg, Mohit Bansal

Detecting customized moments and highlights from videos given natural language (NL) user queries is an important but under-studied topic. One of the challenges in pursuing this direction is the lack of annotated data. To address this issue, we present the Query-based Video Highlights (QVHighlights) dataset. It consists of over 10,000 YouTube videos, covering a wide range of topics, from everyday activities and travel in lifestyle vlog videos to social and political activities in news videos. Each video in the dataset is annotated with: (1) a human-written free-form NL query, (2) relevant moments in the video w.r.t. the query, and (3) five-point scale saliency scores for all query-relevant clips. This comprehensive annotation enables us to develop and evaluate systems that detect relevant moments as well as salient highlights for diverse, flexible user queries. We also present a strong baseline for this task, Moment-DETR, a transformer encoder-decoder model that views moment retrieval as a direct set prediction problem, taking extracted video and query representations as inputs and predicting moment coordinates and saliency scores end-to-end. While our model does not utilize any human prior, we show that it performs competitively when compared to well-engineered architectures. With weakly supervised pretraining using ASR captions, Moment-DETR substantially outperforms previous methods. Lastly, we present several ablations and visualizations of Moment-DETR. Data and code is publicly available at https://github.com/jayleicn/moment_detr.
****

Stochastic optimization under time drift: iterate averaging, step-decay schedules, and high probability guarantees

Joshua Cutler, Dmitriy Drusvyatskiy, Zaid Harchaoui

We consider the problem of minimizing a convex function that is evolving in time

according to unknown and possibly stochastic dynamics. Such problems abound in the machine learning and signal processing literature, under the names of concep t drift and stochastic tracking. We provide novel non-asymptotic convergence gua rantees for stochastic algorithms with iterate averaging, focusing on bounds val id both in expectation and with high probability. Notably, we show that the trac king efficiency of the proximal stochastic gradient method depends only logarith mically on the initialization quality when equipped with a step-decay schedule.
****
Learning Stable Deep Dynamics Models for Partially Observed or Delayed Dynamical Systems
Andreas Schlaginhaufen, Philippe Wenk, Andreas Krause, Florian Dorfler
Learning how complex dynamical systems evolve over time is a key challenge in sy stem identification. For safety critical systems, it is often crucial that the l earned model is guaranteed to converge to some equilibrium point. To this end, n eural ODEs regularized with neural Lyapunov functions are a promising approach w hen states are fully observed. For practical applications however, {\em partial observations} are the norm. As we will demonstrate, initialization of unobserved augmented states can become a key problem for neural ODEs. To alleviate this is sue, we propose to augment the system's state with its history. Inspired by stat e augmentation in discrete-time systems, we thus obtain {\em neural delay differ ential equations}. Based on classical time delay stability analysis, we then sho w how to ensure stability of the learned models, and theoretically analyze our a pproach. Our experiments demonstrate its applicability to stable system identifi cation of partially observed systems and learning a stabilizing feedback policy in delayed feedback control.
****
An Uncertainty Principle is a Price of Privacy-Preserving Microdata
John Abowd, Robert Ashmead, Ryan Cumings-Menon, Simson Garfinkel, Daniel Kifer, Philip Leclerc, William Sexton, Ashley Simpson, Christine Task, Pavel Zhuravlev
Requests for name changes in the electronic proceedings will be accepted with 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 in Ranking under Uncertainty
Ashudeep Singh, David Kempe, Thorsten Joachims
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-auth ors prior to requesting a name change in the electronic proceedings.
****
Generalized Proximal Policy Optimization with Sample Reuse
James Queeney, Yannis Paschalidis, Christos G Cassandras
In real-world decision making tasks, it is critical for data-driven reinforcemen t learning methods to be both stable and sample efficient. On-policy methods typ ically generate reliable policy improvement throughout training, while off-polic y methods make more efficient use of data through sample reuse. In this work, we combine the theoretically supported stability benefits of on-policy algorithms with the sample efficiency of off-policy algorithms. We develop policy improveme nt guarantees that are suitable for the off-policy setting, and connect these bo unds to the clipping mechanism used in Proximal Policy Optimization. This motiva tes an off-policy version of the popular algorithm that we call Generalized Prox imal Policy Optimization with Sample Reuse. We demonstrate both theoretically an d empirically that our algorithm delivers improved performance by effectively ba lancing the competing goals of stability and sample efficiency.
****
Mosaicking to Distill: Knowledge Distillation from Out-of-Domain Data
Gongfan Fang, Yifan Bao, Jie Song, Xinchao Wang, Donglin Xie, Chengchao Shen, Mi ngli Song
Knowledge distillation~(KD) aims to craft a compact student model that imitates

the behavior of a pre-trained teacher in a target domain. Prior KD approaches, d
espite their gratifying results, have largely relied on the premise that \emph{i
n-domain} data is available to carry out the knowledge transfer. Such an assumpt
ion, unfortunately, in many cases violates the practical setting, since the orig
inal training data or even the data domain is often unreachable due to privacy o
r copyright reasons. In this paper, we attempt to tackle an ambitious task, term
ed as \emph{out-of-domain} knowledge distillation~(OOD-KD), which allows us to c
onduct KD using only OOD data that can be readily obtained at a very low cost. A
dmittedly, OOD-KD is by nature a highly challenging task due to the agnostic do
main gap. To this end, we introduce a handy yet surprisingly efficacious approac
h, dubbed as~\textit{MosaicKD}. The key insight behind MosaicKD lies in that, sa
mples from various domains share common local patterns, even though their global
 semantic may vary significantly; these shared local patterns, in turn, can be r
e-assembled analogous to mosaic tiling, to approximate the in-domain data and to
 further alleviating the domain discrepancy. In MosaicKD, this is achieved throu
gh a four-player min-max game, in which a generator, a discriminator, a student
network, are collectively trained in an adversarial manner, partially under the
 guidance of a pre-trained teacher. We validate MosaicKD over {classification an
d semantic segmentation tasks} across various benchmarks, and demonstrate that i
t yields results much superior to the state-of-the-art counterparts on OOD data.
 Our code is available at \url{https://github.com/zju-vipa/MosaicKD}.
****
Batch Active Learning at Scale
Gui Citovsky, Giulia DeSalvo, Claudio Gentile, Lazaros Karydas, Anand Rajagopala
n, Afshin Rostamizadeh, Sanjiv Kumar
The ability to train complex and highly effective models often requires an abund
ance of training data, which can easily become a bottleneck in cost, time, and c
omputational resources. Batch active learning, which adaptively issues batched q
ueries to a labeling oracle, is a common approach for addressing this problem. T
he practical benefits of batch sampling come with the downside of less adaptivit
y and the risk of sampling redundant examples within a batch -- a risk that grow
s with the batch size. In this work, we analyze an efficient active learning alg
orithm, which focuses on the large batch setting. In particular, we show that ou
r sampling method, which combines notions of uncertainty and diversity, easily s
cales to batch sizes (100K-1M) several orders of magnitude larger than used in p
revious studies and provides significant improvements in model training efficien
cy compared to recent baselines. Finally, we provide an initial theoretical anal
ysis, proving label complexity guarantees for a related sampling method, which w
e show is approximately equivalent to our sampling method in specific settings.
****
Joint Semantic Mining for Weakly Supervised RGB-D Salient Object Detection
Jingjing Li, Wei Ji, Qi Bi, Cheng Yan, Miao Zhang, Yongri Piao, Huchuan Lu, Li c
heng
Training saliency detection models with weak supervisions, e.g., image-level tag
s or captions, is appealing as it removes the costly demand of per-pixel annotat
ions. Despite the rapid progress of RGB-D saliency detection in fully-supervised
 setting, it however remains an unexplored territory when only weak supervision
signals are available. This paper is set to tackle the problem of weakly-supervi
sed RGB-D salient object detection. The key insight in this effort is the idea o
f maintaining per-pixel pseudo-labels with iterative refinements by reconciling
the multimodal input signals in our joint semantic mining (JSM). Considering the
 large variations in the raw depth map and the lack of explicit pixel-level supe
rvisions, we propose spatial semantic modeling (SSM) to capture saliency-specifi
c depth cues from the raw depth and produce depth-refined pseudo-labels. Moreove
r, tags and captions are incorporated via a fill-in-the-blank training in our te
xtual semantic modeling (TSM) to estimate the confidences of competing pseudo-la
bels. At test time, our model involves only a light-weight sub-network of the tr
aining pipeline, i.e., it requires only an RGB image as input, thus allowing eff
icient inference. Extensive evaluations demonstrate the effectiveness of our app
roach under the weakly-supervised setting. Importantly, our method could also be

adapted to work in both fully-supervised and unsupervised paradigms. In each of these scenarios, superior performance has been attained by our approach with comparing to the state-of-the-art dedicated methods. As a by-product, a CapS dataset is constructed by augmenting existing benchmark training set with additional image tags and captions.
****
Not All Images are Worth 16x16 Words: Dynamic Transformers for Efficient Image Recognition
Yulin Wang, Rui Huang, Shiji Song, Zeyi Huang, Gao Huang
Vision Transformers (ViT) have achieved remarkable success in large-scale image recognition. They split every 2D image into a fixed number of patches, each of which is treated as a token. Generally, representing an image with more tokens would lead to higher prediction accuracy, while it also results in drastically increased computational cost. To achieve a decent trade-off between accuracy and speed, the number of tokens is empirically set to 16x16 or 14x14. In this paper, we argue that every image has its own characteristics, and ideally the token number should be conditioned on each individual input. In fact, we have observed that there exist a considerable number of “easy” images which can be accurately predicted with a mere number of 4x4 tokens, while only a small fraction of “hard” ones need a finer representation. Inspired by this phenomenon, we propose a Dynamic Transformer to automatically configure a proper number of tokens for each input image. This is achieved by cascading multiple Transformers with increasing numbers of tokens, which are sequentially activated in an adaptive fashion at test time, i.e., the inference is terminated once a sufficiently confident prediction is produced. We further design efficient feature reuse and relationship reuse mechanisms across different components of the Dynamic Transformer to reduce redundant computations. Extensive empirical results on ImageNet, CIFAR-10, and CIFAR-100 demonstrate that our method significantly outperforms the competitive baselines in terms of both theoretical computational efficiency and practical inference speed. Code and pre-trained models (based on PyTorch and MindSpore) are available at https://github.com/blackfeather-wang/Dynamic-Vision-Transformer and https://github.com/blackfeather-wang/Dynamic-Vision-Transformer-MindSpore.
****
Contrastive Learning for Neural Topic Model
Thong Nguyen, Anh Tuan Luu
Recent empirical studies show that adversarial topic models (ATM) can successfully capture semantic patterns of the document by differentiating a document with another dissimilar sample. However, utilizing that discriminative-generative architecture has two important drawbacks: (1) the architecture does not relate similar documents, which has the same document-word distribution of salient words; (2) it restricts the ability to integrate external information, such as sentiments of the document, which has been shown to benefit the training of neural topic model. To address those issues, we revisit the adversarial topic architecture in the view point of mathematical analysis, propose a novel approach to re-formulate discriminative goal as an optimization problem, and design a novel sampling method which facilitates the integration of external variables. The reformulation encourages the model to incorporate the relations among similar samples and enforces the constraint on the similarity among dissimilar ones; while the sampling method, which is based on the internal input and reconstructed output, helps inform the model of salient words contributing to the main topic. Experimental results show that our framework outperforms other state-of-the-art neural topic models in three common benchmark datasets that belong to various domains, vocabulary sizes, and document lengths in terms of topic coherence.
****
Learning in two-player zero-sum partially observable Markov games with perfect recall
Tadashi Kozuno, Pierre Ménard, Remi Munos, Michal Valko
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-auth

ors prior to requesting a name change in the electronic proceedings.
****
A Geometric Structure of Acceleration and Its Role in Making Gradients Small Fast
Jongmin Lee, Chanwoo Park, Ernest Ryu
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
ATISS: Autoregressive Transformers for Indoor Scene Synthesis
Despoina Paschalidou, Amlan Kar, Maria Shugrina, Karsten Kreis, Andreas Geiger, Sanja Fidler
The ability to synthesize realistic and diverse indoor furniture layouts automatically or based on partial input, unlocks many applications, from better interactive 3D tools to data synthesis for training and simulation. In this paper, we present ATISS, a novel autoregressive transformer architecture for creating diverse and plausible synthetic indoor environments, given only the room type and its floor plan. In contrast to prior work, which poses scene synthesis as sequence generation, our model generates rooms as unordered sets of objects. We argue that this formulation is more natural, as it makes ATISS generally useful beyond fully automatic room layout synthesis. For example, the same trained model can be used in interactive applications for general scene completion, partial room re-arrangement with any objects specified by the user, as well as object suggestions for any partial room. To enable this, our model leverages the permutation equivariance of the transformer when conditioning on the partial scene, and is trained to be permutation-invariant across object orderings. Our model is trained end-to-end as an autoregressive generative model using only labeled 3D bounding boxes as supervision. Evaluations on four room types in the 3D-FRONT dataset demonstrate that our model consistently generates plausible room layouts that are more realistic than existing methods.In addition, it has fewer parameters, is simpler to implement and train and runs up to 8 times faster than existing methods.
****
Generalized Depthwise-Separable Convolutions for Adversarially Robust and Efficient Neural Networks
Hassan Dbouk, Naresh Shanbhag
Despite their tremendous successes, convolutional neural networks (CNNs) incur high computational/storage costs and are vulnerable to adversarial perturbations. Recent works on robust model compression address these challenges by combining model compression techniques with adversarial training. But these methods are unable to improve throughput (frames-per-second) on real-life hardware while simultaneously preserving robustness to adversarial perturbations. To overcome this problem, we propose the method of Generalized Depthwise-Separable (GDWS) convolution - an efficient, universal, post-training approximation of a standard 2D convolution. GDWS dramatically improves the throughput of a standard pre-trained network on real-life hardware while preserving its robustness. Lastly, GDWS is scalable to large problem sizes since it operates on pre-trained models and doesn't require any additional training. We establish the optimality of GDWS as a 2D convolution approximator and present exact algorithms for constructing optimal GDWS convolutions under complexity and error constraints. We demonstrate the effectiveness of GDWS via extensive experiments on CIFAR-10, SVHN, and ImageNet datasets. Our code can be found at https://github.com/hsndbk4/GDWS.
****
A Provably Efficient Model-Free Posterior Sampling Method for Episodic Reinforcement Learning
Christoph Dann, Mehryar Mohri, Tong Zhang, Julian Zimmert
Thompson Sampling is one of the most effective methods for contextual bandits and has been generalized to posterior sampling for certain MDP settings. However, existing posterior sampling methods for reinforcement learning are limited by being model-based or lack worst-case theoretical guarantees beyond linear MDPs. Th

is paper proposes a new model-free formulation of posterior sampling that applie
s to more general episodic reinforcement learning problems with theoretical guar
antees. We introduce novel proof techniques to show that under suitable conditio
ns, the worst-case regret of our posterior sampling method matches the best know
n results of optimization based methods. In the linear MDP setting with dimensio
n, the regret of our algorithm scales linearly with the dimension as compared to
 a quadratic dependence of the existing posterior sampling-based exploration alg
orithms.
****
Fast Federated Learning in the Presence of Arbitrary Device Unavailability
Xinran Gu, Kaixuan Huang, Jingzhao Zhang, Longbo Huang
Federated learning (FL) coordinates with numerous heterogeneous devices to colla
boratively train a shared model while preserving user privacy. Despite its multi
ple advantages, FL faces new challenges. One challenge arises when devices drop
out of the training process. In this case, the convergence of popular FL algorit
hms such as FedAvg is severely influenced by the straggling devices. To tackle t
his challenge, we study federated learning algorithms in the presence of arbitra
ry device unavailability and propose an algorithm named Memory-augmented Impatie
nt Federated Averaging (MIFA). Our algorithm efficiently avoids excessive latenc
y induced by inactive devices, and corrects the gradient bias using the memorize
d latest updates from them. We prove that MIFA achieves minimax optimal converge
nce rates on non-i.i.d. data for both strongly convex and non-convex smooth func
tions. We also provide an explicit characterization of the improvement over base
line algorithms through a case study, and validate the results by numerical expe
riments on real-world datasets.
****
On The Structure of Parametric Tournaments with Application to Ranking from Pair
wise Comparisons
Vishnu Veerathu, Arun Rajkumar
Requests for name changes in the electronic proceedings will be accepted with 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.
****
SegFormer: Simple and Efficient Design for Semantic Segmentation with Transforme
rs
Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, Ping Luo
We present SegFormer, a simple, efficient yet powerful semantic segmentation fra
mework which unifies Transformers with lightweight multilayer perceptron (MLP) d
ecoders. SegFormer has two appealing features: 1) SegFormer comprises a novel hi
erarchically structured Transformer encoder which outputs multiscale features. I
t does not need positional encoding, thereby avoiding the interpolation of posit
ional codes which leads to decreased performance when the testing resolution dif
fers from training. 2) SegFormer avoids complex decoders. The proposed MLP decod
er aggregates information from different layers, and thus combining both local a
ttention and global attention to render powerful representations. We show that t
his simple and lightweight design is the key to efficient segmentation on Transf
ormers. We scale our approach up to obtain a series of models from SegFormer-B0
 to Segformer-B5, which reaches much better performance and efficiency than prev
ious counterparts.For example, SegFormer-B4 achieves 50.3% mIoU on ADE20K with 6
4M parameters, being 5x smaller and 2.2% better than the previous best method. O
ur best model, SegFormer-B5, achieves 84.0% mIoU on Cityscapes validation set an
d shows excellent zero-shot robustness on Cityscapes-C.
****
Fairness via Representation Neutralization
Mengnan Du, Subhabrata Mukherjee, Guanchu Wang, Ruixiang Tang, Ahmed Awadallah,
Xia Hu
Existing bias mitigation methods for DNN models primarily work on learning debia
sed encoders. This process not only requires a lot of instance-level annotations
 for sensitive attributes, it also does not guarantee that all fairness sensitiv

e information has been removed from the encoder. To address these limitations, we explore the following research question: Can we reduce the discrimination of DNN models by only debiasing the classification head, even with biased representations as inputs? To this end, we propose a new mitigation technique, namely, Representation Neutralization for Fairness (RNF) that achieves fairness by debiasing only the task-specific classification head of DNN models. To this end, we leverage samples with the same ground-truth label but different sensitive attributes, and use their neutralized representations to train the classification head of the DNN model. The key idea of RNF is to discourage the classification head from capturing spurious correlation between fairness sensitive information in encoder representations with specific class labels. To address low-resource settings with no access to sensitive attribute annotations, we leverage a bias-amplified model to generate proxy annotations for sensitive attributes. Experimental results over several benchmark datasets demonstrate our RNF framework to effectively reduce discrimination of DNN models with minimal degradation in task-specific performance.
****
Residual Relaxation for Multi-view Representation Learning
Yifei Wang, Zhengyang Geng, Feng Jiang, Chuming Li, Yisen Wang, Jiansheng Yang, Zhouchen Lin
Multi-view methods learn representations by aligning multiple views of the same image and their performance largely depends on the choice of data augmentation. In this paper, we notice that some other useful augmentations, such as image rotation, are harmful for multi-view methods because they cause a semantic shift that is too large to be aligned well. This observation motivates us to relax the exact alignment objective to better cultivate stronger augmentations. Taking image rotation as a case study, we develop a generic approach, Pretext-aware Residual Relaxation (Prelax), that relaxes the exact alignment by allowing an adaptive residual vector between different views and encoding the semantic shift through pretext-aware learning. Extensive experiments on different backbones show that our method can not only improve multi-view methods with existing augmentations, but also benefit from stronger image augmentations like rotation.
****
Do Vision Transformers See Like Convolutional Neural Networks?
Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, Alexey Dosovitskiy
Convolutional neural networks (CNNs) have so far been the de-facto model for visual data. Recent work has shown that (Vision) Transformer models (ViT) can achieve comparable or even superior performance on image classification tasks. This raises a central question: how are Vision Transformers solving these tasks? Are they acting like convolutional networks, or learning entirely different visual representations? Analyzing the internal representation structure of ViTs and CNNs on image classification benchmarks, we find striking differences between the two architectures, such as ViT having more uniform representations across all layers. We explore how these differences arise, finding crucial roles played by self-attention, which enables early aggregation of global information, and ViT residual connections, which strongly propagate features from lower to higher layers. We study the ramifications for spatial localization, demonstrating ViTs successfully preserve input spatial information, with noticeable effects from different classification methods. Finally, we study the effect of (pretraining) dataset scale on intermediate features and transfer learning, and conclude with a discussion on connections to new architectures such as the MLP-Mixer.
****
Optimization-Based Algebraic Multigrid Coarsening Using Reinforcement Learning
Ali Taghibakhshi, Scott MacLachlan, Luke Olson, Matthew West
Large sparse linear systems of equations are ubiquitous in science and engineering, such as those arising from discretizations of partial differential equations. Algebraic multigrid (AMG) methods are one of the most common methods of solving such linear systems, with an extensive body of underlying mathematical theory. A system of linear equations defines a graph on the set of unknowns and each le

vel of a multigrid solver requires the selection of an appropriate coarse graph along with restriction and interpolation operators that map to and from the coarse representation. The efficiency of the multigrid solver depends critically on this selection and many selection methods have been developed over the years. Recently, it has been demonstrated that it is possible to directly learn the AMG interpolation and restriction operators, given a coarse graph selection. In this paper, we consider the complementary problem of learning to coarsen graphs for a multigrid solver, a necessary step in developing fully learnable AMG methods. We propose a method using a reinforcement learning (RL) agent based on graph neural networks (GNNs), which can learn to perform graph coarsening on small planar training graphs and then be applied to unstructured large planar graphs, assuming bounded node degree. We demonstrate that this method can produce better coarse graphs than existing algorithms, even as the graph size increases and other properties of the graph are varied. We also propose an efficient inference procedure for performing graph coarsening that results in linear time complexity in graph size.
****
Delayed Propagation Transformer: A Universal Computation Engine towards Practical Control in Cyber-Physical Systems
Wenqing Zheng, Qiangqiang Guo, Hao Yang, Peihao Wang, Zhangyang Wang
Multi-agent control is a central theme in the Cyber-Physical Systems (CPS). However, current control methods either receive non-Markovian states due to insufficient sensing and decentralized design, or suffer from poor convergence. This paper presents the Delayed Propagation Transformer (DePT), a new transformer-based model that specializes in the global modeling of CPS while taking into account the immutable constraints from the physical world. DePT induces a cone-shaped spatial-temporal attention prior, which injects the information propagation and aggregation principles and enables a global view. With physical constraint inductive bias baked into its design, our DePT is ready to plug and play for a broad class of multi-agent systems. The experimental results on one of the most challenging CPS -- network-scale traffic signal control system in the open world -- show that our model outperformed the state-of-the-art expert methods on synthetic and real-world datasets. Our codes are released at: https://github.com/VITA-Group/DePT.
****
Explaining Latent Representations with a Corpus of Examples
Jonathan Crabbe, Zhaozhi Qian, Fergus Imrie, Mihaela van der Schaar
Modern machine learning models are complicated. Most of them rely on convoluted latent representations of their input to issue a prediction. To achieve greater transparency than a black-box that connects inputs to predictions, it is necessary to gain a deeper understanding of these latent representations. To that aim, we propose SimplEx: a user-centred method that provides example-based explanations with reference to a freely selected set of examples, called the corpus. SimplEx uses the corpus to improve the user’s understanding of the latent space with post-hoc explanations answering two questions: (1) Which corpus examples explain the prediction issued for a given test example? (2) What features of these corpus examples are relevant for the model to relate them to the test example? SimplEx provides an answer by reconstructing the test latent representation as a mixture of corpus latent representations. Further, we propose a novel approach, the integrated Jacobian, that allows SimplEx to make explicit the contribution of each corpus feature in the mixture. Through experiments on tasks ranging from mortality prediction to image classification, we demonstrate that these decompositions are robust and accurate. With illustrative use cases in medicine, we show that SimplEx empowers the user by highlighting relevant patterns in the corpus that explain model representations. Moreover, we demonstrate how the freedom in choosing the corpus allows the user to have personalized explanations in terms of examples that are meaningful for them.
****
Explaining heterogeneity in medial entorhinal cortex with task-driven neural networks

Aran Nayebi, Alexander Attinger, Malcolm Campbell, Kiah Hardcastle, Isabel Low, Caitlin S Mallory, Gabriel Mel, Ben Sorscher, Alex H Williams, Surya Ganguli, Lisa Giocomo, Dan Yamins
Medial entorhinal cortex (MEC) supports a wide range of navigational and memory related behaviors.Well-known experimental results have revealed specialized cell types in MEC --- e.g. grid, border, and head-direction cells --- whose highly stereotypical response profiles are suggestive of the role they might play in supporting MEC functionality. However, the majority of MEC neurons do not exhibit stereotypical firing patterns.How should the response profiles of these more "heterogeneous" cells be described, and how do they contribute to behavior?In this work, we took a computational approach to addressing these questions.We first performed a statistical analysis that shows that heterogeneous MEC cells are just as reliable in their response patterns as the more stereotypical cell types, suggesting that they have a coherent functional role.Next, we evaluated a spectrum of candidate models in terms of their ability to describe the response profiles of both stereotypical and heterogeneous MEC cells.We found that recently developed task-optimized neural network models are substantially better than traditional grid cell-centric models at matching most MEC neuronal response profiles --- including those of grid cells themselves --- despite not being explicitly trained for this purpose.Specific choices of network architecture (such as gated nonlinearities and an explicit intermediate place cell representation) have an important effect on the ability of the model to generalize to novel scenarios, with the best of these models closely approaching the noise ceiling of the data itself.We then performed in silico experiments on this model to address questions involving the relative functional relevance of various cell types, finding that heterogeneous cells are likely to be just as involved in downstream functional outcomes (such as path integration) as grid and border cells.Finally, inspired by recent data showing that, going beyond their spatial response selectivity, MEC cells are also responsive to non-spatial rewards, we introduce a new MEC model that performs reward-modulated path integration.We find that this unified model matches neural recordings across all variable-reward conditions.Taken together, our results point toward a conceptually principled goal-driven modeling approach for moving future experimental and computational efforts beyond overly-simplistic single-cell stereotypes.
****
Beyond Smoothness: Incorporating Low-Rank Analysis into Nonparametric Density Estimation
Robert A. Vandermeulen, Antoine Ledent
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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-View Representation Learning via Total Correlation Objective
HyeongJoo Hwang, Geon-Hyeong Kim, Seunghoon Hong, Kee-Eung Kim
Multi-View Representation Learning (MVRL) aims to discover a shared representation of observations from different views with the complex underlying correlation. In this paper, we propose a variational approach which casts MVRL as maximizing the amount of total correlation reduced by the representation, aiming to learn a shared latent representation that is informative yet succinct to capture the correlation among multiple views. To this end, we introduce a tractable surrogate objective function under the proposed framework, which allows our method to fuse and calibrate the observations in the representation space. From the information-theoretic perspective, we show that our framework subsumes existing multi-view generative models. Lastly, we show that our approach straightforwardly extends to the Partial MVRL (PMVRL) setting, where the observations are missing without any regular pattern. We demonstrate the effectiveness of our approach in the multi-view translation and classification tasks, outperforming strong baseline methods.
****

FACMAC: Factored Multi-Agent Centralised Policy Gradients
Bei Peng, Tabish Rashid, Christian Schroeder de Witt, Pierre-Alexandre Kamienny, Philip Torr, Wendelin Boehmer, 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-authors prior to requesting a name change in the electronic proceedings.
****
EDGE: Explaining Deep Reinforcement Learning Policies
Wenbo Guo, Xian Wu, Usmann Khan, Xinyu Xing
With the rapid development of deep reinforcement learning (DRL) techniques, there is an increasing need to understand and interpret DRL policies. While recent research has developed explanation methods to interpret how an agent determines its moves, they cannot capture the importance of actions/states to a game's final result. In this work, we propose a novel self-explainable model that augments a Gaussian process with a customized kernel function and an interpretable predictor. Together with the proposed model, we also develop a parameter learning procedure that leverages inducing points and variational inference to improve learning efficiency. Using our proposed model, we can predict an agent's final rewards from its game episodes and extract time step importance within episodes as strategy-level explanations for that agent. Through experiments on Atari and MuJoCo games, we verify the explanation fidelity of our method and demonstrate how to employ interpretation to understand agent behavior, discover policy vulnerabilities, remediate policy errors, and even defend against adversarial attacks.
****
Learning to Assimilate in Chaotic Dynamical Systems
Michael McCabe, Jed Brown
The accuracy of simulation-based forecasting in chaotic systems is heavily dependent on high-quality estimates of the system state at the beginning of the forecast. Data assimilation methods are used to infer these initial conditions by systematically combining noisy, incomplete observations and numerical models of system dynamics to produce highly effective estimation schemes. We introduce a self-supervised framework, which we call \textit{amortized assimilation}, for learning to assimilate in dynamical systems. Amortized assimilation combines deep learning-based denoising with differentiable simulation, using independent neural networks to assimilate specific observation types while connecting the gradient flow between these sub-tasks with differentiable simulation and shared recurrent memory. This hybrid architecture admits a self-supervised training objective which is minimized by an unbiased estimator of the true system state even in the presence of only noisy training data. Numerical experiments across several chaotic benchmark systems highlight the improved effectiveness of our approach compared to widely-used data assimilation methods.
****
Object-aware Contrastive Learning for Debiased Scene Representation
Sangwoo Mo, Hyunwoo Kang, Kihyuk Sohn, Chun-Liang Li, Jinwoo Shin
Contrastive self-supervised learning has shown impressive results in learning visual representations from unlabeled images by enforcing invariance against different data augmentations. However, the learned representations are often contextually biased to the spurious scene correlations of different objects or object and background, which may harm their generalization on the downstream tasks. To tackle the issue, we develop a novel object-aware contrastive learning framework that first (a) localizes objects in a self-supervised manner and then (b) debias scene correlations via appropriate data augmentations considering the inferred object locations. For (a), we propose the contrastive class activation map (ContraCAM), which finds the most discriminative regions (e.g., objects) in the image compared to the other images using the contrastively trained models. We further improve the ContraCAM to detect multiple objects and entire shapes via an iterative refinement procedure. For (b), we introduce two data augmentations based on ContraCAM, object-aware random crop and background mixup, which reduce contextual and background biases during contrastive self-supervised learning, respectivel

y. Our experiments demonstrate the effectiveness of our representation learning framework, particularly when trained under multi-object images or evaluated unde r the background (and distribution) shifted images. Code is available at https://github.com/alinlab/object-aware-contrastive.
****
Evaluating Efficient Performance Estimators of Neural Architectures
Xuefei Ning, Changcheng Tang, Wenshuo Li, Zixuan Zhou, Shuang Liang, Huazhong Yang, Yu Wang
Conducting efficient performance estimations of neural architectures is a major challenge in neural architecture search (NAS). To reduce the architecture training costs in NAS, one-shot estimators (OSEs) amortize the architecture training costs by sharing the parameters of one supernet between all architectures. Recently, zero-shot estimators (ZSEs) that involve no training are proposed to further reduce the architecture evaluation cost. Despite the high efficiency of these estimators, the quality of such estimations has not been thoroughly studied. In this paper, we conduct an extensive and organized assessment of OSEs and ZSEs on five NAS benchmarks: NAS-Bench-101/201/301, and NDS ResNet/ResNeXt-A. Specifically, we employ a set of NAS-oriented criteria to study the behavior of OSEs and ZSEs, and reveal their biases and variances. After analyzing how and why the OSE estimations are unsatisfying, we explore how to mitigate the correlation gap of OSEs from three perspectives. Through our analysis, we give out suggestions for future application and development of efficient architecture performance estimators. Furthermore, the analysis framework proposed in our work could be utilized in future research to give a more comprehensive understanding of newly designed architecture performance estimators. The code is available at https://github.com/walkerning/aw_nas.
****
A-NeRF: Articulated Neural Radiance Fields for Learning Human Shape, Appearance, and Pose
Shih-Yang Su, Frank Yu, Michael Zollhoefer, Helge Rhodin
While deep learning reshaped the classical motion capture pipeline with feed-forward networks, generative models are required to recover fine alignment via iterative refinement. Unfortunately, the existing models are usually hand-crafted or learned in controlled conditions, only applicable to limited domains. We propose a method to learn a generative neural body model from unlabelled monocular videos by extending Neural Radiance Fields (NeRFs). We equip them with a skeleton to apply to time-varying and articulated motion. A key insight is that implicit models require the inverse of the forward kinematics used in explicit surface models. Our reparameterization defines spatial latent variables relative to the pose of body parts and thereby overcomes ill-posed inverse operations with an overparameterization. This enables learning volumetric body shape and appearance from scratch while jointly refining the articulated pose; all without ground truth labels for appearance, pose, or 3D shape on the input videos. When used for novel-view-synthesis and motion capture, our neural model improves accuracy on diverse datasets.
****
Differential Privacy Over Riemannian Manifolds
Matthew Reimherr, Karthik Bharath, Carlos Soto
In this work we consider the problem of releasing a differentially private statistical summary that resides on a Riemannian manifold. We present an extension of the Laplace or K-norm mechanism that utilizes intrinsic distances and volumes on the manifold. We also consider in detail the specific case where the summary is the Fr\'echet mean of data residing on a manifold. We demonstrate that our mechanism is rate optimal and depends only on the dimension of the manifold, not on the dimension of any ambient space, while also showing how ignoring the manifold structure can decrease the utility of the sanitized summary. We illustrate our framework in two examples of particular interest in statistics: the space of symmetric positive definite matrices, which is used for covariance matrices, and the sphere, which can be used as a space for modeling discrete distributions.

****
How can classical multidimensional scaling go wrong?
Rishi Sonthalia, Greg Van Buskirk, Benjamin Raichel, 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.
****
Modeling Heterogeneous Hierarchies with Relation-specific Hyperbolic Cones
Yushi Bai, Zhitao Ying, Hongyu Ren, Jure Leskovec
Hierarchical relations are prevalent and indispensable for organizing human knowledge captured by a knowledge graph (KG). The key property of hierarchical relations is that they induce a partial ordering over the entities, which needs to be modeled in order to allow for hierarchical reasoning. However, current KG embeddings can model only a single global hierarchy (single global partial ordering) and fail to model multiple heterogeneous hierarchies that exist in a single KG. Here we present ConE (Cone Embedding), a KG embedding model that is able to simultaneously model multiple hierarchical as well as non-hierarchical relations in a knowledge graph. ConE embeds entities into hyperbolic cones and models relations as transformations between the cones. In particular, ConE uses cone containment constraints in different subspaces of the hyperbolic embedding space to capture multiple heterogeneous hierarchies. Experiments on standard knowledge graph benchmarks show that ConE obtains state-of-the-art performance on hierarchical reasoning tasks as well as knowledge graph completion task on hierarchical graphs. In particular, our approach yields new state-of-the-art Hits@1 of 45.3% on WN18RR and 16.1% on DDB14 (0.231 MRR). As for hierarchical reasoning task, our approach outperforms previous best results by an average of 20% across the three datasets.
****
Non-asymptotic Error Bounds for Bidirectional GANs
Shiao Liu, Yunfei Yang, Jian Huang, Yuling Jiao, Yang Wang
We derive nearly sharp bounds for the bidirectional GAN (BiGAN) estimation error under the Dudley distance between the latent joint distribution and the data joint distribution with appropriately specified architecture of the neural networks used in the model. To the best of our knowledge, this is the first theoretical guarantee for the bidirectional GAN learning approach. An appealing feature of our results is that they do not assume the reference and the data distributions to have the same dimensions or these distributions to have bounded support. These assumptions are commonly assumed in the existing convergence analysis of the unidirectional GANs but may not be satisfied in practice. Our results are also applicable to the Wasserstein bidirectional GAN if the target distribution is assumed to have a bounded support. To prove these results, we construct neural network functions that push forward an empirical distribution to another arbitrary empirical distribution on a possibly different-dimensional space. We also develop a novel decomposition of the integral probability metric for the error analysis of bidirectional GANs. These basic theoretical results are of independent interest and can be applied to other related learning problems.
****
Confidence-Aware Imitation Learning from Demonstrations with Varying Optimality
Songyuan Zhang, ZHANGJIE CAO, Dorsa Sadigh, Yanan Sui
Most existing imitation learning approaches assume the demonstrations are drawn from experts who are optimal, but relaxing this assumption enables us to use a wider range of data. Standard imitation learning may learn a suboptimal policy from demonstrations with varying optimality. Prior works use confidence scores or rankings to capture beneficial information from demonstrations with varying optimality, but they suffer from many limitations, e.g., manually annotated confidence scores or high average optimality of demonstrations. In this paper, we propose a general framework to learn from demonstrations with varying optimality that jointly learns the confidence score and a well-performing policy. Our approach, Confidence-Aware Imitation Learning (CAIL) learns a well-performing policy from

confidence-reweighted demonstrations, while using an outer loss to track the per
formance of our model and to learn the confidence. We provide theoretical guaran
tees on the convergence of CAIL and evaluate its performance in both simulated a
nd real robot experiments.Our results show that CAIL significantly outperforms o
ther imitation learning methods from demonstrations with varying optimality. We
further show that even without access to any optimal demonstrations, CAIL can st
ill learn a successful policy, and outperforms prior work.
****
Answering Complex Causal Queries With the Maximum Causal Set Effect
Zachary Markovich
The standard tools of causal inference have been developed to answer simple caus
al queries which can be easily formalized as a small number of statistical estim
ands in the context of a particular structural causal model (SCM); however, scie
ntific theories often make diffuse predictions about a large number of causal va
riables. This article proposes a framework for parameterizing such complex causa
l queries as the maximum difference in causal effects associated with two sets o
f causal variables that have a researcher specified probability of occurring. We
 term this estimand the Maximum Causal Set Effect (MCSE) and develop an estimato
r for it that is asymptotically consistent and conservative in finite samples un
der assumptions that are standard in the causal inference literature. This estim
ator is also asymptotically normal and amenable to the non-parametric bootstrap,
 facilitating classical statistical inference about this novel estimand. We comp
are this estimator to more common latent variable approaches and find that it ca
n uncover larger causal effects in both real world and simulated data.
****
Identifiability in inverse reinforcement learning
Haoyang Cao, Samuel Cohen, Lukasz Szpruch
Inverse reinforcement learning attempts to reconstruct the reward function in a
Markov decision problem, using observations of agent actions. As already observe
d in Russell [1998] the problem is ill-posed, and the reward function is not ide
ntifiable, even under the presence of perfect information about optimal behavior
. We provide a resolution to this non-identifiability for problems with entropy
regularization. For a given environment, we fully characterize the reward functi
ons leading to a given policy and demonstrate that, given demonstrations of acti
ons for the same reward under two distinct discount factors, or under sufficient
ly different environments, the unobserved reward can be recovered up to a consta
nt. We also give general necessary and sufficient conditions for reconstruction
of time-homogeneous rewards on finite horizons, and for action-independent rewar
ds, generalizing recent results of Kim et al. [2021] and Fu et al. [2018].
****
A Probabilistic State Space Model for Joint Inference from Differential Equation
s and Data
Jonathan Schmidt, Nicholas Krämer, Philipp Hennig
Mechanistic models with differential equations are a key component of scientific
 applications of machine learning. Inference in such models is usually computati
onally demanding because it involves repeatedly solving the differential equatio
n. The main problem here is that the numerical solver is hard to combine with st
andard inference techniques. Recent work in probabilistic numerics has developed
 a new class of solvers for ordinary differential equations (ODEs) that phrase t
he solution process directly in terms of Bayesian filtering. We here show that t
his allows such methods to be combined very directly, with conceptual and numeri
cal ease, with latent force models in the ODE itself. It then becomes possible t
o perform approximate Bayesian inference on the latent force as well as the ODE
solution in a single, linear complexity pass of an extended Kalman filter / smoo
ther — that is, at the cost of computing a single ODE solution. We demonstrate t
he expressiveness and performance of the algorithm by training, among others, a
non-parametric SIRD model on data from the COVID-19 outbreak.
****
On Plasticity, Invariance, and Mutually Frozen Weights in Sequential Task Learni
ng

Julian Zilly, Alessandro Achille, Andrea Censi, Emilio Frazzoli
Plastic neural networks have the ability to adapt to new tasks. However, in a co ntinual learning setting, the configuration of parameters learned in previous ta sks can severely reduce the adaptability to future tasks. In particular, we show that, when using weight decay, weights in successive layers of a deep network m ay become "mutually frozen". This has a double effect: on the one hand, it makes the network updates more invariant to nuisance factors, providing a useful bias for future tasks. On the other hand, it can prevent the network from learning n ew tasks that require significantly different features. In this context, we find that the local input sensitivity of a deep model is correlated with its ability to adapt, thus leading to an intriguing trade-off between adaptability and inva riance when training a deep model more than once. We then show that a simple int ervention that "resets" the mutually frozen connections can improve transfer lea rning on a variety of visual classification tasks. The efficacy of "resetting" i tself depends on the size of the target dataset and the difference of the pre-tr aining and target domains, allowing us to achieve state-of-the-art results on so me datasets.
****
Provably Efficient Black-Box Action Poisoning Attacks Against Reinforcement Lear ning
Guanlin Liu, Lifeng LAI
Due to the broad range of applications of reinforcement learning (RL), understan ding the effects of adversarial attacks against RL model is essential for the sa fe applications of this model. Prior theoretical works on adversarial attacks ag ainst RL mainly focus on either reward poisoning attacks or environment poisonin g attacks. In this paper, we introduce a new class of attacks named action poiso ning attacks, where an adversary can change the action signal selected by the ag ent. Compared with existing attack models, the attacker’s ability in the propose d action poisoning attack model is more restricted, which brings some design cha llenges. We study the action poisoning attack in both white-box and black-box se ttings. We introduce an adaptive attack scheme called LCB-H, which works for mos t RL agents in the black-box setting. We prove that LCB-H attack can force any e fficient RL agent, whose dynamic regret scales sublinearly with the total number of steps taken, to choose actions according to a policy selected by the attacke r very frequently, with only sublinear cost. In addition, we apply LCB-H attack against a very popular model-free RL algorithm: UCB-H. We show that, even in bla ck-box setting, by spending only logarithm cost, the proposed LCB-H attack schem e can force the UCB-H agent to choose actions according to the policy selected b y the attacker very frequently.
****
Fast Approximation of the Sliced-Wasserstein Distance Using Concentration of Ran dom Projections
Kimia Nadjahi, Alain Durmus, Pierre E Jacob, Roland Badeau, Umut Simsekli
The Sliced-Wasserstein distance (SW) is being increasingly used in machine learn ing applications as an alternative to the Wasserstein distance and offers signif icant computational and statistical benefits. Since it is defined as an expectat ion over random projections, SW is commonly approximated by Monte Carlo. We adop t a new perspective to approximate SW by making use of the concentration of meas ure phenomenon: under mild assumptions, one-dimensional projections of a high-di mensional random vector are approximately Gaussian. Based on this observation, w e develop a simple deterministic approximation for SW. Our method does not requi re sampling a number of random projections, and is therefore both accurate and e asy to use compared to the usual Monte Carlo approximation. We derive nonasympto tical guarantees for our approach, and show that the approximation error goes to zero as the dimension increases, under a weak dependence condition on the data distribution. We validate our theoretical findings on synthetic datasets, and il lustrate the proposed approximation on a generative modeling problem.
****
Causal Navigation by Continuous-time Neural Networks
Charles Vorbach, Ramin Hasani, Alexander Amini, Mathias Lechner, Daniela Rus

Imitation learning enables high-fidelity, vision-based learning of policies with
in rich, photorealistic environments. However, such techniques often rely on tra
ditional discrete-time neural models and face difficulties in generalizing to do
main shifts by failing to account for the causal relationships between the agent
 and the environment. In this paper, we propose a theoretical and experimental f
ramework for learning causal representations using continuous-time neural networ
ks, specifically over their discrete-time counterparts. We evaluate our method i
n the context of visual-control learning of drones over a series of complex task
s, ranging from short- and long-term navigation, to chasing static and dynamic o
bjects through photorealistic environments. Our results demonstrate that causal
continuous-time deep models can perform robust navigation tasks, where advanced
recurrent models fail. These models learn complex causal control representations
 directly from raw visual inputs and scale to solve a variety of tasks using imi
tation learning.
****
Global Convergence of Online Optimization for Nonlinear Model Predictive Control
Sen Na
We study a real-time iteration (RTI) scheme for solving online optimization prob
lem appeared in nonlinear optimal control. The proposed RTI scheme modifies the
existing RTI-based model predictive control (MPC) algorithm, by selecting the st
epsize of each Newton step at each sampling time using a differentiable exact au
gmented Lagrangian. The scheme can adaptively select the penalty parameters of a
ugmented Lagrangian on the fly, which are shown to be stabilized after certain t
ime periods. We prove under generic assumptions that, by involving stepsize sele
ction instead of always using a full Newton step (like what most of the existing
 RTIs do), the scheme converges globally: for any initial point, the KKT residua
ls of the subproblems converge to zero. A key step is to show that augmented Lag
rangian keeps decreasing as horizon moves forward. We demonstrate the global con
vergence behavior of the proposed RTI scheme in a numerical experiment.
****
Argmax Flows and Multinomial Diffusion: Learning Categorical Distributions
Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, Max Welling
Generative flows and diffusion models have been predominantly trained on ordinal
 data, for example natural images. This paper introduces two extensions of flows
 and diffusion for categorical data such as language or image segmentation: Argm
ax Flows and Multinomial Diffusion. Argmax Flows are defined by a composition of
 a continuous distribution (such as a normalizing flow), and an argmax function.
 To optimize this model, we learn a probabilistic inverse for the argmax that li
fts the categorical data to a continuous space. Multinomial Diffusion gradually
adds categorical noise in a diffusion process, for which the generative denoisin
g process is learned. We demonstrate that our method outperforms existing dequan
tization approaches on text modelling and modelling on image segmentation maps i
n log-likelihood.
****
Learning with User-Level Privacy
Daniel Levy, Ziteng Sun, Kareem Amin, Satyen Kale, Alex Kulesza, Mehryar Mohri,
Ananda Theertha Suresh
Requests for name changes in the electronic proceedings will be accepted with 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.
****
Don’t Generate Me: Training Differentially Private Generative Models with Sinkho
rn Divergence
Tianshi Cao, Alex Bie, Arash Vahdat, Sanja Fidler, Karsten Kreis
Although machine learning models trained on massive data have led to breakthroug
hs in several areas, their deployment in privacy-sensitive domains remains limit
ed due to restricted access to data. Generative models trained with privacy cons
traints on private data can sidestep this challenge, providing indirect access t
o private data instead. We propose DP-Sinkhorn, a novel optimal transport-based

generative method for learning data distributions from private data with differe
ntial privacy. DP-Sinkhorn minimizes the Sinkhorn divergence, a computationally
efficient approximation to the exact optimal transport distance, between the mod
el and data in a differentially private manner and uses a novel technique for co
ntrolling the bias-variance trade-off of gradient estimates. Unlike existing app
roaches for training differentially private generative models, which are mostly
based on generative adversarial networks, we do not rely on adversarial objectiv
es, which are notoriously difficult to optimize, especially in the presence of n
oise imposed by privacy constraints. Hence, DP-Sinkhorn is easy to train and dep
loy. Experimentally, we improve upon the state-of-the-art on multiple image mode
ling benchmarks and show differentially private synthesis of informative RGB ima
ges.
****
Keeping Your Eye on the Ball: Trajectory Attention in Video Transformers
Mandela Patrick, Dylan Campbell, Yuki Asano, Ishan Misra, Florian Metze, Christo
ph Feichtenhofer, Andrea Vedaldi, João F. Henriques
Requests for name changes in the electronic proceedings will be accepted with 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 Bayesian Optimistic Sampling
Brendan O'Donoghue, Tor Lattimore
Requests for name changes in the electronic proceedings will be accepted with no
 questions asked. However name changes may cause bibliographic tracking issues.
  Authors are asked to consider this carefully and discuss it with their co-auth
ors prior to requesting a name change in the electronic proceedings.
****
Cross-modal Domain Adaptation for Cost-Efficient Visual Reinforcement Learning
Xiong-Hui Chen, Shengyi Jiang, Feng Xu, Zongzhang Zhang, Yang Yu
In visual-input sim-to-real scenarios, to overcome the reality gap between image
s rendered in simulators and those from the real world, domain adaptation, i.e.,
 learning an aligned representation space between simulators and the real world,
 then training and deploying policies in the aligned representation, is a promis
ing direction. Previous methods focus on same-modal domain adaptation. However,
those methods require building and running simulators that render high-quality i
mages, which can be difficult and costly. In this paper, we consider a more cost
-efficient setting of visual-input sim-to-real where only low-dimensional states
 are simulated. We first point out that the objective of learning mapping functi
ons in previous methods that align the representation spaces is ill-posed, prone
 to yield an incorrect mapping. When the mapping crosses modalities, previous me
thods are easier to fail. Our algorithm, Cross-mOdal Domain Adaptation with Sequ
ential structure (CODAS), mitigates the ill-posedness by utilizing the sequentia
l nature of the data sampling process in RL tasks. Experiments on MuJoCo and Han
d Manipulation Suite tasks show that the agents deployed with our method achieve
 similar performance as it has in the source domain, while those deployed with p
revious methods designed for same-modal domain adaptation suffer a larger perfor
mance gap.
****
D2C: Diffusion-Decoding Models for Few-Shot Conditional Generation
Abhishek Sinha, Jiaming Song, Chenlin Meng, Stefano Ermon
Conditional generative models of high-dimensional images have many applications,
 but supervision signals from conditions to images can be expensive to acquire.
This paper describes Diffusion-Decoding models with Contrastive representations
(D2C), a paradigm for training unconditional variational autoencoders (VAE) for
few-shot conditional image generation. D2C uses a learned diffusion-based prior
over the latent representations to improve generation and contrastive self-super
vised learning to improve representation quality. D2C can adapt to novel generat
ion tasks, conditioned on labels or manipulation constraints, by learning from a
s few as 100 labeled examples. On conditional generation from new labels, D2C ac

hieves superior performance over state-of-the-art VAEs and diffusion models. On conditional image manipulation, D2C generations are two orders of magnitude faster to produce over StyleGAN2 ones and are preferred by 50% - 60% of the human evaluators in a double-blind study. We release our code at https://github.com/jiamings/d2c.
****
Continual Auxiliary Task Learning
Matthew McLeod, Chunlok Lo, Matthew Schlegel, Andrew Jacobsen, Raksha Kumaraswamy, Martha White, Adam White
Learning auxiliary tasks, such as multiple predictions about the world, can provide many benefits to reinforcement learning systems. A variety of off-policy learning algorithms have been developed to learn such predictions, but as yet there is little work on how to adapt the behavior to gather useful data for those off-policy predictions. In this work, we investigate a reinforcement learning system designed to learn a collection of auxiliary tasks, with a behavior policy learning to take actions to improve those auxiliary predictions. We highlight the inherent non-stationarity in this continual auxiliary task learning problem, for both prediction learners and the behavior learner. We develop an algorithm based on successor features that facilitates tracking under non-stationary rewards, and prove the separation into learning successor features and rewards provides convergence rate improvements. We conduct an in-depth study into the resulting multi-prediction learning system.
****
Constrained Two-step Look-Ahead Bayesian Optimization
Yunxiang Zhang, Xiangyu Zhang, Peter Frazier
Recent advances in computationally efficient non-myopic Bayesian optimization offer improved query efficiency over traditional myopic methods like expected improvement, with only a modest increase in computational cost. These advances have been largely limited to unconstrained BO methods with only a few exceptions which require heavy computation. For instance, one existing multi-step lookahead constrained BO method (Lam & Willcox, 2017) relies on computationally expensive unreliable brute force derivative-free optimization of a Monte Carlo rollout acquisition function. Methods that use the reparameterization trick for more efficient derivative-based optimization of non-myopic acquisition functions in the unconstrained setting, like sample average approximation and infinitesimal perturbation analysis, do not extend: constraints introduce discontinuities in the sampled acquisition function surface. Moreover, we argue here that being non-myopic is even more important in constrained problems because fear of violating constraints pushes myopic methods away from sampling the boundary between feasible and infeasible regions, slowing the discovery of optimal solutions with tight constraints. In this paper, we propose a computationally efficient two-step lookahead constrained Bayesian optimization acquisition function (2-OPT-C) supporting both sequential and batch settings. To enable fast acquisition function optimization, we develop a novel likelihood ratio-based unbiased estimator of the gradient of the two-step optimal acquisition function that does not use the reparameterization trick. In numerical experiments, 2-OPT-C typically improves query efficiency by 2x or more over previous methods, and in some cases by 10x or more.
****
Learning with Labeling Induced Abstentions
Kareem Amin, Giulia DeSalvo, Afshin Rostamizadeh
Consider a setting where we wish to automate an expensive task with a machine learning algorithm using a limited labeling resource. In such settings, examples routed for labeling are often out of scope for the machine learning algorithm. For example, in a spam detection setting, human reviewers not only provide labeled data but are such high-quality detectors of spam that examples routed to them no longer require machine evaluation. As a consequence, the distribution of examples routed to the machine is intimately tied to the process generating labels. We introduce a formalization of this setting, and give an algorithm that simultaneously learns a model and decides when to request a label by leveraging ideas from both the abstention and active learning literatures. We prove an upper bound

on the algorithm's label complexity and a matching lower bound for any algorithm in this setting. We conduct a thorough set of experiments including an ablation study to test different components of our algorithm. We demonstrate the effectiveness of an efficient version of our algorithm over margin sampling on a variety of datasets.
****
SQALER: Scaling Question Answering by Decoupling Multi-Hop and Logical Reasoning
Mattia Atzeni, Jasmina Bogojeska, Andreas Loukas
State-of-the-art approaches to reasoning and question answering over knowledge graphs (KGs) usually scale with the number of edges and can only be applied effectively on small instance-dependent subgraphs. In this paper, we address this issue by showing that multi-hop and more complex logical reasoning can be accomplished separately without losing expressive power. Motivated by this insight, we propose an approach to multi-hop reasoning that scales linearly with the number of relation types in the graph, which is usually significantly smaller than the number of edges or nodes. This produces a set of candidate solutions that can be provably refined to recover the solution to the original problem. Our experiments on knowledge-based question answering show that our approach solves the multi-hop MetaQA dataset, achieves a new state-of-the-art on the more challenging WebQuestionsSP, is orders of magnitude more scalable than competitive approaches, and can achieve compositional generalization out of the training distribution.
****
Out-of-Distribution Generalization in Kernel Regression
Abdulkadir Canatar, Blake Bordelon, Cengiz Pehlevan
In real word applications, data generating process for training a machine learning model often differs from what the model encounters in the test stage. Understanding how and whether machine learning models generalize under such distributional shifts have been a theoretical challenge. Here, we study generalization in kernel regression when the training and test distributions are different using methods from statistical physics. Using the replica method, we derive an analytical formula for the out-of-distribution generalization error applicable to any kernel and real datasets. We identify an overlap matrix that quantifies the mismatch between distributions for a given kernel as a key determinant of generalization performance under distribution shift. Using our analytical expressions we elucidate various generalization phenomena including possible improvement in generalization when there is a mismatch. We develop procedures for optimizing training and test distributions for a given data budget to find best and worst case generalizations under the shift. We present applications of our theory to real and synthetic datasets and for many kernels. We compare results of our theory applied to Neural Tangent Kernel with simulations of wide networks and show agreement. We analyze linear regression in further depth.
****
FL-WBC: Enhancing Robustness against Model Poisoning Attacks in Federated Learning from a Client Perspective
Jingwei Sun, Ang Li, Louis DiValentin, Amin Hassanzadeh, Yiran Chen, Hai Li
Federated learning (FL) is a popular distributed learning framework that trains a global model through iterative communications between a central server and edge devices. Recent works have demonstrated that FL is vulnerable to model poisoning attacks. Several server-based defense approaches (e.g. robust aggregation), have been proposed to mitigate such attacks. However, we empirically show that under extremely strong attacks, these defensive methods fail to guarantee the robustness of FL. More importantly, we observe that as long as the global model is polluted, the impact of attacks on the global model will remain in subsequent rounds even if there are no subsequent attacks. In this work, we propose a client-based defense, named White Blood Cell for Federated Learning (FL-WBC), which can mitigate model poisoning attacks that have already polluted the global model. The key idea of FL-WBC is to identify the parameter space where long-lasting attack effect on parameters resides and perturb that space during local training. Furthermore, we derive a certified robustness guarantee against model poisoning attacks and a convergence guarantee to FedAvg after applying our FL-WBC. We conduct

experiments on FasionMNIST and CIFAR10 to evaluate the defense against state-of-the-art model poisoning attacks. The results demonstrate that our method can effectively mitigate model poisoning attack impact on the global model within 5 communication rounds with nearly no accuracy drop under both IID and Non-IID settings. Our defense is also complementary to existing server-based robust aggregation approaches and can further improve the robustness of FL under extremely strong attacks.
****

Chebyshev-Cantelli PAC-Bayes-Bennett Inequality for the Weighted Majority Vote
Yi-Shan Wu, Andres Masegosa, Stephan Lorenzen, Christian Igel, Yevgeny Seldin
We present a new second-order oracle bound for the expected risk of a weighted majority vote. The bound is based on a novel parametric form of the Chebyshev-Cantelli inequality (a.k.a. one-sided Chebyshev's), which is amenable to efficient minimization. The new form resolves the optimization challenge faced by prior oracle bounds based on the Chebyshev-Cantelli inequality, the C-bounds [Germain et al., 2015], and, at the same time, it improves on the oracle bound based on second order Markov's inequality introduced by Masegosa et al. [2020]. We also derive a new concentration of measure inequality, which we name PAC-Bayes-Bennett, since it combines PAC-Bayesian bounding with Bennett's inequality. We use it for empirical estimation of the oracle bound. The PAC-Bayes-Bennett inequality improves on the PAC-Bayes-Bernstein inequality of Seldin et al. [2012]. We provide an empirical evaluation demonstrating that the new bounds can improve on the work of Masegosa et al. [2020]. Both the parametric form of the Chebyshev-Cantelli inequality and the PAC-Bayes-Bennett inequality may be of independent interest for the study of concentration of measure in other domains.
****

A Multi-Implicit Neural Representation for Fonts
Pradyumna Reddy, Zhifei Zhang, Zhaowen Wang, Matthew Fisher, Hailin Jin, Niloy Mitra
Fonts are ubiquitous across documents and come in a variety of styles. They are either represented in a native vector format or rasterized to produce fixed resolution images. In the first case, the non-standard representation prevents benefiting from latest network architectures for neural representations; while, in the latter case, the rasterized representation, when encoded via networks, results in loss of data fidelity, as font-specific discontinuities like edges and corners are difficult to represent using neural networks. Based on the observation that complex fonts can be represented by a superposition of a set of simpler occupancy functions, we introduce multi-implicits to represent fonts as a permutation-invariant set of learned implict functions, without losing features (e.g., edges and corners). However, while multi-implicits locally preserve font features, obtaining supervision in the form of ground truth multi-channel signals is a problem in itself. Instead, we propose how to train such a representation with only local supervision, while the proposed neural architecture directly finds globally consistent multi-implicits for font families. We extensively evaluate the proposed representation for various tasks including reconstruction, interpolation, and synthesis to demonstrate clear advantages with existing alternatives. Additionally, the representation naturally enables glyph completion, wherein a single characteristic font is used to synthesize a whole font family in the target style.
****

OctField: Hierarchical Implicit Functions for 3D Modeling
Jia-Heng Tang, Weikai Chen, jie Yang, Bo Wang, Songrun Liu, Bo Yang, Lin Gao
Recent advances in localized implicit functions have enabled neural implicit representation to be scalable to large scenes.However, the regular subdivision of 3D space employed by these approaches fails to take into account the sparsity of the surface occupancy and the varying granularities of geometric details. As a result, its memory footprint grows cubically with the input volume, leading to a prohibitive computational cost even at a moderately dense decomposition. In this work, we present a learnable hierarchical implicit representation for 3D surfaces, coded OctField, that allows high-precision encoding of intricate surfaces wi

th low memory and computational budget. The key to our approach is an adaptive decomposition of 3D scenes that only distributes local implicit functions around the surface of interest. We achieve this goal by introducing a hierarchical octree structure to adaptively subdivide the 3D space according to the surface occupancy and the richness of part geometry. As octree is discrete and non-differentiable, we further propose a novel hierarchical network that models the subdivision of octree cells as a probabilistic process and recursively encodes and decodes both octree structure and surface geometry in a differentiable manner. We demonstrate the value of OctField for a range of shape modeling and reconstruction tasks, showing superiority over alternative approaches.
****

The Inductive Bias of Quantum Kernels
Jonas Kübler, Simon Buchholz, Bernhard Schölkopf
It has been hypothesized that quantum computers may lend themselves well to applications in machine learning. In the present work, we analyze function classes defined via quantum kernels. Quantum computers offer the possibility to efficiently compute inner products of exponentially large density operators that are classically hard to compute. However, having an exponentially large feature space renders the problem of generalization hard. Furthermore, being able to evaluate inner products in high dimensional spaces efficiently by itself does not guarantee a quantum advantage, as already classically tractable kernels can correspond to high- or infinite-dimensional reproducing kernel Hilbert spaces (RKHS).   We analyze the spectral properties of quantum kernels and find that we can expect an advantage if their RKHS is low dimensional and contains functions that are hard to compute classically. If the target function is known to lie in this class, this implies a quantum advantage, as the quantum computer can encode this inductive bias, whereas there is no classically efficient way to constrain the function class in the same way. However, we show that finding suitable quantum kernels is not easy because the kernel evaluation might require exponentially many measurements.   In conclusion, our message is a somewhat sobering one: we conjecture that quantum machine learning models can offer speed-ups only if we manage to encode knowledge about the problem at hand into quantum circuits, while encoding the same bias into a classical model would be hard. These situations may plausibly occur when learning on data generated by a quantum process, however, they appear to be harder to come by for classical datasets.
****

An Exponential Improvement on the Memorization Capacity of Deep Threshold Networks
Shashank Rajput, Kartik Sreenivasan, Dimitris Papailiopoulos, 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.
****

Pretraining Representations for Data-Efficient Reinforcement Learning
Max Schwarzer, Nitarshan Rajkumar, Michael Noukhovitch, Ankesh Anand, Laurent Charlin, R Devon Hjelm, Philip Bachman, Aaron C. Courville
Data efficiency is a key challenge for deep reinforcement learning. We address this problem by using unlabeled data to pretrain an encoder which is then finetuned on a small amount of task-specific data. To encourage learning representations which capture diverse aspects of the underlying MDP, we employ a combination of latent dynamics modelling and unsupervised goal-conditioned RL. When limited to 100k steps of interaction on Atari games (equivalent to two hours of human experience), our approach significantly surpasses prior work combining offline representation pretraining with task-specific finetuning, and compares favourably with other pretraining methods that require orders of magnitude more data. Our approach shows particular promise when combined with larger models as well as more diverse, task-aligned observational data -- approaching human-level performance and data-efficiency on Atari in our best setting.
****

Universal Approximation Using Well-Conditioned Normalizing Flows
Holden Lee, Chirag Pabbaraju, Anish Prasad Sevekari, Andrej Risteski
Normalizing flows are a widely used class of latent-variable generative models with a tractable likelihood. Affine-coupling models [Dinh et al., 2014, 2016] are a particularly common type of normalizing flows, for which the Jacobian of the latent-to-observable-variable transformation is triangular, allowing the likelihood to be computed in linear time. Despite the widespread usage of affine couplings, the special structure of the architecture makes understanding their representational power challenging. The question of universal approximation was only recently resolved by three parallel papers [Huang et al., 2020, Zhang et al., 2020, Koehler et al., 2020] – who showed reasonably regular distributions can be approximated arbitrarily well using affine couplings – albeit with networks with a nearly-singular Jacobian. As ill-conditioned Jacobians are an obstacle for likelihood-based training, the fundamental question remains: which distributions can be approximated using well-conditioned affine coupling flows? In this paper, we show that any log-concave distribution can be approximated using well-conditioned affine-coupling flows. In terms of proof techniques, we uncover and leverage deep connections between affine coupling architectures, underdamped Langevin dynamics (a stochastic differential equation often used to sample from Gibbs measures) and Hénon maps (a structured dynamical system that appears in the study of symplectic diffeomorphisms). In terms of informing practice, we approximate a padded version of the input distribution with iid Gaussians – a strategy which Koehler et al. [2020] empirically observed to result in better-conditioned flows, but had hitherto no theoretical grounding. Our proof can thus be seen as providing theoretical evidence for the benefits of Gaussian padding when training normalizing flows.
****
On the Validity of Modeling SGD with Stochastic Differential Equations (SDEs)
Zhiyuan Li, Sadhika Malladi, Sanjeev Arora
It is generally recognized that finite learning rate (LR), in contrast to infinitesimal LR, is important for good generalization in real-life deep nets. Most attempted explanations propose approximating finite-LR SGD with Itô Stochastic Differential Equations (SDEs), but formal justification for this approximation (e.g., Li et al., 2019) only applies to SGD with tiny LR. Experimental verification of the approximation appears computationally infeasible. The current paper clarifies the picture with the following contributions: (a) An efficient simulation algorithm SVAG that provably converges to the conventionally used Itô SDE approximation. (b) A theoretically motivated testable necessary condition for the SDE approximation and its most famous implication, the linear scaling rule (Goyal et al., 2017), to hold.(c) Experiments using this simulation to demonstrate that the previously proposed SDE approximation can meaningfully capture the training and generalization properties of common deep nets.
****
Proportional Participatory Budgeting with Additive Utilities
Dominik Peters, Grzegorz Pierczy■ski, Piotr Skowron
We study voting rules for participatory budgeting, where a group of voters collectively decides which projects should be funded using a common budget. We allow the projects to have arbitrary costs, and the voters to have arbitrary additive valuations over the projects. We formulate two axioms that guarantee proportional representation to groups of voters with common interests. To the best of our knowledge, all known rules for participatory budgeting do not satisfy either of the two axioms; in addition we show that the most prominent proportional rule for committee elections, Proportional Approval Voting, cannot be adapted to arbitrary costs nor to additive valuations so that it would satisfy our axioms of proportionality. We construct a simple and attractive voting rule that satisfies one of our axioms (for arbitrary costs and arbitrary additive valuations), and that can be evaluated in polynomial time. We prove that our other stronger axiom is also satisfiable, though by a computationally more expensive and less natural voting rule.
****

Disentangling the Roles of Curation, Data-Augmentation and the Prior in the Cold Posterior Effect
Lorenzo Noci, Kevin Roth, Gregor Bachmann, Sebastian Nowozin, Thomas Hofmann
The “cold posterior effect” (CPE) in Bayesian deep learning describes the disturbing observation that the predictive performance of Bayesian neural networks can be significantly improved if the Bayes posterior is artificially sharpened using a temperature parameter T <1. The CPE is problematic in theory and practice and since the effect was identified many researchers have proposed hypotheses to explain the phenomenon. However, despite this intensive research effort the effect remains poorly understood. In this work we provide novel and nuanced evidence relevant to existing explanations for the cold posterior effect, disentangling three hypotheses: 1. The dataset curation hypothesis of Aitchison (2020): we show empirically that the CPE does not arise in a real curated data set but can be produced in a controlled experiment with varying curation strength. 2. The data augmentation hypothesis of Izmailov et al. (2021) and Fortuin et al. (2021): we show empirically that data augmentation is sufficient but not necessary for the CPE to be present. 3. The bad prior hypothesis of Wenzel et al. (2020): we use a simple experiment evaluating the relative importance of the prior and the likelihood, strongly linking the CPE to the prior. Our results demonstrate how the CPE can arise in isolation from synthetic curation, data augmentation, and bad priors. Cold posteriors observed “in the wild” are therefore unlikely to arise from a single simple cause; as a result, we do not expect a simple “fix” for cold posteriors.
****
Sanity Checks for Lottery Tickets: Does Your Winning Ticket Really Win the Jackpot?
Xiaolong Ma, Geng Yuan, Xuan Shen, Tianlong Chen, Xuxi Chen, Xiaohan Chen, Ning Liu, Minghai Qin, Sijia Liu, Zhangyang Wang, Yanzhi Wang
There have been long-standing controversies and inconsistencies over the experiment setup and criteria for identifying the "winning ticket" in literature. To reconcile such, we revisit the definition of lottery ticket hypothesis, with comprehensive and more rigorous conditions. Under our new definition, we show concrete evidence to clarify whether the winning ticket exists across the major DNN architectures and/or applications. Through extensive experiments, we perform quantitative analysis on the correlations between winning tickets and various experimental factors, and empirically study the patterns of our observations. We find that the key training hyperparameters, such as learning rate and training epochs, as well as the architecture characteristics such as capacities and residual connections, are all highly correlated with whether and when the winning tickets can be identified. Based on our analysis, we summarize a guideline for parameter settings in regards of specific architecture characteristics, which we hope to catalyze the research progress on the topic of lottery ticket hypothesis. Our codes are publicly available at: https://github.com/boone891214/sanity-check-LTH.
****
Collaborative Causal Discovery with Atomic Interventions
Raghavendra Addanki, Shiva Kasiviswanathan
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Towards optimally abstaining from prediction with OOD test examples
Adam Kalai, Varun Kanade
A common challenge across all areas of machine learning is that training data is not distributed like test data, due to natural shifts or adversarial examples; such examples are referred to as out-of-distribution (OOD) test examples. We consider a model where one may abstain from predicting, at a fixed cost. In particular, our transductive abstention algorithm takes labeled training examples and unlabeled test examples as input, and provides predictions with optimal prediction loss guarantees. The loss bounds match standard generalization bounds when tes

t examples are i.i.d. from the training distribution, but add an additional term
 that is the cost of abstaining times the statistical distance between the train
 and test distribution (or the fraction of adversarial examples). For linear reg
ression, we give a polynomial-time algorithm based on Celis-Dennis-Tapia optimiz
ation algorithms. For binary classification, we show how to efficiently implemen
t it using a proper agnostic learner (i.e., an Empirical Risk Minimizer) for the
 class of interest. Our work builds on recent work of Goldwasser, Kalais, and Mo
ntasser (2020) who gave error and abstention guarantees for transductive binary
classification.
****
TokenLearner: Adaptive Space-Time Tokenization for Videos
Michael Ryoo, AJ Piergiovanni, Anurag Arnab, Mostafa Dehghani, Anelia Angelova
In this paper, we introduce a novel visual representation learning which relies
on a handful of adaptively learned tokens, and which is applicable to both image
 and video understanding tasks. Instead of relying on hand-designed splitting st
rategies to obtain visual tokens and processing a large number of densely sample
d patches for attention, our approach learns to mine important tokens in visual
data. This results in efficiently and effectively finding a few important visual
 tokens and enables modeling of pairwise attention between such tokens, over a l
onger temporal horizon for videos, or the spatial content in image frames. Our e
xperiments demonstrate strong performance on several challenging benchmarks for
video recognition tasks. Importantly, due to our tokens being adaptive, we accom
plish competitive results at significantly reduced computational cost. We establ
ish new state-of-the-arts on multiple video datasets, including Kinetics-400, Ki
netics-600, Charades, and AViD.
****
Learning in Multi-Stage Decentralized Matching Markets
Xiaowu Dai, Michael Jordan
Matching markets are often organized in a multi-stage and decentralized manner.
 Moreover, participants in real-world matching markets often have uncertain pref
erences. This article develops a framework for learning optimal strategies in su
ch settings, based on a nonparametric statistical approach and variational analy
sis. We propose an efficient algorithm, built upon concepts of "lower uncertain
ty bound" and "calibrated decentralized matching," for maximizing the participan
ts' expected payoff. We show that there exists a welfare-versus-fairness trade-o
ff that is characterized by the uncertainty level of acceptance. Participants w
ill strategically act in favor of a low uncertainty level to reduce competition
and increase expected payoff. We prove that participants can be better off with
multi-stage matching compared to single-stage matching. We demonstrate aspects o
f the theoretical predictions through simulations and an experiment using real d
ata from college admissions.
****
Non-asymptotic convergence bounds for Wasserstein approximation using point clou
ds
Quentin Mérigot, Filippo Santambrogio, Clément SARRAZIN
Requests for name changes in the electronic proceedings will be accepted with 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 Interlocking Dynamics of Cooperative Rationalization
Mo Yu, Yang Zhang, Shiyu Chang, Tommi Jaakkola
Selective rationalization explains the prediction of complex neural networks by
finding a small subset of the input that is sufficient to predict the neural mod
el output. The selection mechanism is commonly integrated into the model itself
by specifying a two-component cascaded system consisting of a rationale generato
r, which makes a binary selection of the input features (which is the rationale)
, and a predictor, which predicts the output based only on the selected features
. The components are trained jointly to optimize prediction performance. In this
 paper, we reveal a major problem with such cooperative rationalization paradigm

--- model interlocking. Inter-locking arises when the predictor overfits to the features selected by the generator thus reinforcing the generator's selection even if the selected rationales are sub-optimal. The fundamental cause of the interlocking problem is that the rationalization objective to be minimized is concave with respect to the generator’s selection policy. We propose a new rationalization framework, called A2R, which introduces a third component into the architecture, a predictor driven by soft attention as opposed to selection. The generator now realizes both soft and hard attention over the features and these are fed into the two different predictors. While the generator still seeks to support the original predictor performance, it also minimizes a gap between the two predictors. As we will show theoretically, since the attention-based predictor exhibits a better convexity property, A2R can overcome the concavity barrier. Our experiments on two synthetic benchmarks and two real datasets demonstrate that A2R can significantly alleviate the interlock problem and find explanations that better align with human judgments.
****

Adversarial Robustness without Adversarial Training: A Teacher-Guided Curriculum Learning Approach
Anindya Sarkar, Anirban Sarkar, Sowrya Gali, Vineeth N Balasubramanian
Current SOTA adversarially robust models are mostly based on adversarial training (AT) and differ only by some regularizers either at inner maximization or outer minimization steps. Being repetitive in nature during the inner maximization step, they take a huge time to train. We propose a non-iterative method that enforces the following ideas during training. Attribution maps are more aligned to the actual object in the image for adversarially robust models compared to naturally trained models. Also, the allowed set of pixels to perturb an image (that changes model decision) should be restricted to the object pixels only, which reduces the attack strength by limiting the attack space. Our method achieves significant performance gains with a little extra effort (10-20%) over existing AT models and outperforms all other methods in terms of adversarial as well as natural accuracy. We have performed extensive experimentation with CIFAR-10, CIFAR-100, and TinyImageNet datasets and reported results against many popular strong adversarial attacks to prove the effectiveness of our method.
****

Tactical Optimism and Pessimism for Deep Reinforcement Learning
Ted Moskovitz, Jack Parker-Holder, Aldo Pacchiano, Michael Arbel, Michael Jordan
In recent years, deep off-policy actor-critic algorithms have become a dominant approach to reinforcement learning for continuous control. One of the primary drivers of this improved performance is the use of pessimistic value updates to address function approximation errors, which previously led to disappointing performance. However, a direct consequence of pessimism is reduced exploration, running counter to theoretical support for the efficacy of optimism in the face of uncertainty. So which approach is best? In this work, we show that the most effective degree of optimism can vary both across tasks and over the course of learning. Inspired by this insight, we introduce a novel deep actor-critic framework, Tactical Optimistic and Pessimistic (TOP) estimation, which switches between optimistic and pessimistic value learning online. This is achieved by formulating the selection as a multi-arm bandit problem. We show in a series of continuous control tasks that TOP outperforms existing methods which rely on a fixed degree of optimism, setting a new state of the art in challenging pixel-based environments. Since our changes are simple to implement, we believe these insights can easily be incorporated into a multitude of off-policy algorithms.
****

Towards Hyperparameter-free Policy Selection for Offline Reinforcement Learning
Siyuan Zhang, Nan 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.
****

FjORD: Fair and Accurate Federated Learning under heterogeneous targets with Ord
ered Dropout
Samuel Horváth, Stefanos Laskaridis, Mario Almeida, Ilias Leontiadis, Stylianos
Venieris, Nicholas Lane
Federated Learning (FL) has been gaining significant traction across different M
L tasks, ranging from vision to keyboard predictions. In large-scale deployments
, client heterogeneity is a fact and constitutes a primary problem for fairness,
 training performance and accuracy. Although significant efforts have been made
into tackling statistical data heterogeneity, the diversity in the processing ca
pabilities and network bandwidth of clients, termed system heterogeneity, has re
mained largely unexplored. Current solutions either disregard a large portion of
 available devices or set a uniform limit on the model's capacity, restricted by
 the least capable participants.In this work, we introduce Ordered Dropout, a me
chanism that achieves an ordered, nested representation of knowledge in Neural N
etworks and enables the extraction of lower footprint submodels without the need
 for retraining. We further show that for linear maps our Ordered Dropout is equ
ivalent to SVD. We employ this technique, along with a self-distillation method
ology, in the realm of FL in a framework called FjORD. FjORD alleviates the prob
lem of client system heterogeneity by tailoring the model width to the client's
capabilities. Extensive evaluation on both CNNs and RNNs across diverse modaliti
es shows that FjORD consistently leads to significant performance gains over sta
te-of-the-art baselines while maintaining its nested structure.
****
Optimal Uniform OPE and Model-based Offline Reinforcement Learning in Time-Homog
eneous, Reward-Free and Task-Agnostic Settings
Ming Yin, 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.
****
MixSeq: Connecting Macroscopic Time Series Forecasting with Microscopic Time Ser
ies Data
Zhibo Zhu, Ziqi Liu, Ge Jin, Zhiqiang Zhang, Lei Chen, Jun Zhou, Jianyong Zhou
Time series forecasting is widely used in business intelligence, e.g., forecast
stock market price, sales, and help the analysis of data trend. Most time series
 of interest are macroscopic time series that are aggregated from microscopic da
ta. However, instead of directly modeling the macroscopic time series, rare lite
rature studied the forecasting of macroscopic time series by leveraging data on
the microscopic level. In this paper, we assume that the microscopic time series
 follow some unknown mixture probabilistic distributions. We theoretically show
that as we identify the ground truth latent mixture components, the estimation o
f time series from each component could be improved because of lower variance, t
hus benefitting the estimation of macroscopic time series as well. Inspired by t
he power of Seq2seq and its variants on the modeling of time series data, we pro
pose Mixture of Seq2seq (MixSeq), an end2end mixture model to cluster microscopi
c time series, where all the components come from a family of Seq2seq models par
ameterized by different parameters. Extensive experiments on both synthetic and
real-world data show the superiority of our approach.
****
Pareto Domain Adaptation
fangrui lv, Jian Liang, Kaixiong Gong, Shuang Li, Chi Harold Liu, Han Li, Di Liu
, Guoren Wang
Domain adaptation (DA) attempts to transfer the knowledge from a labeled source
domain to an unlabeled target domain that follows different distribution from th
e source. To achieve this, DA methods include a source classification objective
to extract the source knowledge and a domain alignment objective to diminish the
 domain shift, ensuring knowledge transfer. Typically, former DA methods adopt s
ome weight hyper-parameters to linearly combine the training objectives to form
an overall objective. However, the gradient directions of these objectives may c

onflict with each other due to domain shift. Under such circumstances, the linea
r optimization scheme might decrease the overall objective value at the expense
of damaging one of the training objectives, leading to restricted solutions. In
this paper, we rethink the optimization scheme for DA from a gradient-based pers
pective. We propose a Pareto Domain Adaptation (ParetoDA) approach to control th
e overall optimization direction, aiming to cooperatively optimize all training
objectives. Specifically, to reach a desirable solution on the target domain, we
 design a surrogate loss mimicking target classification. To improve target-pred
iction accuracy to support the mimicking, we propose a target-prediction refinin
g mechanism which exploits domain labels via Bayes' theorem. On the other hand,
since prior knowledge of weighting schemes for objectives is often unavailable t
o guide optimization to approach the optimal solution on the target domain, we p
ropose a dynamic preference mechanism to dynamically guide our cooperative optim
ization by the gradient of the surrogate loss on a held-out unlabeled target dat
aset. Our theoretical analyses show that the held-out data can guide but will no
t be over-fitted by the optimization. Extensive experiments on image classificat
ion and semantic segmentation benchmarks demonstrate the effectiveness of Pareto
DA
****
Divergence Frontiers for Generative Models: Sample Complexity, Quantization Effe
cts, and Frontier Integrals
Lang Liu, Krishna Pillutla, Sean Welleck, Sewoong Oh, Yejin Choi, Zaid Harchaoui
The spectacular success of deep generative models calls for quantitative tools t
o measure their statistical performance. Divergence frontiers have recently been
 proposed as an evaluation framework for generative models, due to their ability
 to measure the quality-diversity trade-off inherent to deep generative modeling
. We establish non-asymptotic bounds on the sample complexity of divergence fron
tiers. We also introduce frontier integrals which provide summary statistics of
divergence frontiers. We show how smoothed estimators such as Good-Turing or Kri
chevsky-Trofimov can overcome the missing mass problem and lead to faster rates
of convergence. We illustrate the theoretical results with numerical examples fr
om natural language processing and computer vision.
****
Consistency Regularization for Variational Auto-Encoders
Samarth Sinha, Adji Bousso Dieng
Variational Auto-Encoders (VAEs) are a powerful approach to unsupervised learnin
g. They enable scalable approximate posterior inference in latent-variable model
s using variational inference. A VAE posits a variational family parameterized b
y a deep neural network---called an encoder---that takes data as input. This enc
oder is shared across all the observations, which amortizes the cost of inferenc
e. However the encoder of a VAE has the undesirable property that it maps a give
n observation and a semantics-preserving transformation of it to different laten
t representations. This "inconsistency" of the encoder lowers the quality of the
 learned representations, especially for downstream tasks, and also negatively a
ffects generalization. In this paper, we propose a regularization method to enfo
rce consistency in VAEs. The idea is to minimize the Kullback-Leibler (KL) diver
gence between the variational distribution when conditioning on the observation
and the variational distribution when conditioning on a random semantics-preserv
ing transformation of this observation. This regularization is applicable to any
 VAE. In our experiments we apply it to four different VAE variants on several b
enchmark datasets and found it always improves the quality of the learned repres
entations but also leads to better generalization. In particular, when applied t
o the Nouveau VAE (NVAE), our regularization method yields state-of-the-art perf
ormance on MNIST, CIFAR-10, and CELEBA. We also applied our method to 3D data an
d found it learns representations of superior quality as measured by accuracy on
 a downstream classification task. Finally, we show our method can even outperfo
rm the triplet loss, an advanced and popular contrastive learning-based method f
or representation learning.
****
Score-based Generative Neural Networks for Large-Scale Optimal Transport

Max Daniels, Tyler Maunu, Paul Hand
We consider the fundamental problem of sampling the optimal transport coupling between given source and target distributions. In certain cases, the optimal transport plan takes the form of a one-to-one mapping from the source support to the target support, but learning or even approximating such a map is computationally challenging for large and high-dimensional datasets due to the high cost of linear programming routines and an intrinsic curse of dimensionality. We study instead the Sinkhorn problem, a regularized form of optimal transport whose solutions are couplings between the source and the target distribution. We introduce a novel framework for learning the Sinkhorn coupling between two distributions in the form of a score-based generative model. Conditioned on source data, our procedure iterates Langevin Dynamics to sample target data according to the regularized optimal coupling. Key to this approach is a neural network parametrization of the Sinkhorn problem, and we prove convergence of gradient descent with respect to network parameters in this formulation. We demonstrate its empirical success on a variety of large scale optimal transport tasks.
****

Interactive Label Cleaning with Example-based Explanations
Stefano Teso, Andrea Bontempelli, Fausto Giunchiglia, Andrea Passerini
We tackle sequential learning under label noise in applications where a human supervisor can be queried to relabel suspicious examples. Existing approaches are flawed, in that they only relabel incoming examples that look "suspicious" to the model. As a consequence, those mislabeled examples that elude (or don't undergo) this cleaning step end up tainting the training data and the model with no further chance of being cleaned. We propose CINCER, a novel approach that cleans both new and past data by identifying \emph{pairs of mutually incompatible examples}. Whenever it detects a suspicious example, CINCER identifies a counter-example in the training set that - according to the model - is maximally incompatible with the suspicious example, and asks the annotator to relabel either or both examples, resolving this possible inconsistency. The counter-examples are chosen to be maximally incompatible, so to serve as \emph{explanations} of the model's suspicion, and highly influential, so to convey as much information as possible if relabeled. CINCER achieves this by leveraging an efficient and robust approximation of influence functions based on the Fisher information matrix (FIM). Our extensive empirical evaluation shows that clarifying the reasons behind the model's suspicions by cleaning the counter-examples helps in acquiring substantially better data and models, especially when paired with our FIM approximation.
****

Gradient Descent on Two-layer Nets: Margin Maximization and Simplicity Bias
Kaifeng Lyu, Zhiyuan Li, Runzhe Wang, Sanjeev Arora
The generalization mystery of overparametrized deep nets has motivated efforts to understand how gradient descent (GD) converges to low-loss solutions that generalize well. Real-life neural networks are initialized from small random values and trained with cross-entropy loss for classification (unlike the "lazy" or "NTK" regime of training where analysis was more successful), and a recent sequence of results (Lyu and Li, 2020; Chizat and Bach, 2020; Ji and Telgarsky, 2020) provide theoretical evidence that GD may converge to the "max-margin" solution with zero loss, which presumably generalizes well. However, the global optimality of margin is proved only in some settings where neural nets are infinitely or exponentially wide. The current paper is able to establish this global optimality for two-layer Leaky ReLU nets trained with gradient flow on linearly separable and symmetric data, regardless of the width. The analysis also gives some theoretical justification for recent empirical findings (Kalimeris et al., 2019) on the so-called simplicity bias of GD towards linear or other "simple" classes of solutions, especially early in training. On the pessimistic side, the paper suggests that such results are fragile. A simple data manipulation can make gradient flow converge to a linear classifier with suboptimal margin.
****

Glance-and-Gaze Vision Transformer
Qihang Yu, Yingda Xia, Yutong Bai, Yongyi Lu, Alan L. Yuille, Wei Shen

Recently, there emerges a series of vision Transformers, which show superior per formance with a more compact model size than conventional convolutional neural n etworks, thanks to the strong ability of Transformers to model long-range depend encies. However, the advantages of vision Transformers also come with a price: S elf-attention, the core part of Transformer, has a quadratic complexity to the i nput sequence length. This leads to a dramatic increase of computation and memor y cost with the increase of sequence length, thus introducing difficulties when applying Transformers to the vision tasks that require dense predictions based o n high-resolution feature maps.In this paper, we propose a new vision Transforme r, named Glance-and-Gaze Transformer (GG-Transformer), to address the aforementi oned issues. It is motivated by the Glance and Gaze behavior of human beings whe n recognizing objects in natural scenes, with the ability to efficiently model b oth long-range dependencies and local context. In GG-Transformer, the Glance and Gaze behavior is realized by two parallel branches: The Glance branch is achiev ed by performing self-attention on the adaptively-dilated partitions of the inpu t, which leads to a linear complexity while still enjoying a global receptive fi eld; The Gaze branch is implemented by a simple depth-wise convolutional layer, which compensates local image context to the features obtained by the Glance mec hanism. We empirically demonstrate our method achieves consistently superior per formance over previous state-of-the-art Transformers on various vision tasks and benchmarks.
****

Stochastic $L^\natural$-convex Function Minimization

Haixiang Zhang, Zeyu Zheng, Javad Lavaei

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

Self-Supervised GANs with Label Augmentation

Liang Hou, Huawei Shen, Qi Cao, Xueqi Cheng

Recently, transformation-based self-supervised learning has been applied to gene rative adversarial networks (GANs) to mitigate catastrophic forgetting in the di scriminator by introducing a stationary learning environment. However, the separ ate self-supervised tasks in existing self-supervised GANs cause a goal inconsis tent with generative modeling due to the fact that their self-supervised classif iers are agnostic to the generator distribution. To address this problem, we pro pose a novel self-supervised GAN that unifies the GAN task with the self-supervi sed task by augmenting the GAN labels (real or fake) via self-supervision of dat a transformation. Specifically, the original discriminator and self-supervised c lassifier are unified into a label-augmented discriminator that predicts the aug mented labels to be aware of both the generator distribution and the data distri bution under every transformation, and then provide the discrepancy between them to optimize the generator. Theoretically, we prove that the optimal generator c ould converge to replicate the real data distribution. Empirically, we show that the proposed method significantly outperforms previous self-supervised and data augmentation GANs on both generative modeling and representation learning acros s benchmark datasets.
****

Shape As Points: A Differentiable Poisson Solver

Songyou Peng, Chiyu Jiang, Yiyi Liao, Michael Niemeyer, Marc Pollefeys, Andreas Geiger

In recent years, neural implicit representations gained popularity in 3D reconst ruction due to their expressiveness and flexibility. However, the implicit natur e of neural implicit representations results in slow inference times and require s careful initialization. In this paper, we revisit the classic yet ubiquitous p oint cloud representation and introduce a differentiable point-to-mesh layer usi ng a differentiable formulation of Poisson Surface Reconstruction (PSR) which al lows for a GPU-accelerated fast solution of the indicator function given an orie nted point cloud. The differentiable PSR layer allows us to efficiently and diff

erentiably bridge the explicit 3D point representation with the 3D mesh via the implicit indicator field, enabling end-to-end optimization of surface reconstruc tion metrics such as Chamfer distance. This duality between points and meshes he nce allows us to represent shapes as oriented point clouds, which are explicit, lightweight and expressive. Compared to neural implicit representations, our Sha pe-As-Points (SAP) model is more interpretable, lightweight, and accelerates inf erence time by one order of magnitude. Compared to other explicit representation s such as points, patches, and meshes, SAP produces topology-agnostic, watertigh t manifold surfaces. We demonstrate the effectiveness of SAP on the task of surf ace reconstruction from unoriented point clouds and learning-based reconstructio n.
****

Outcome-Driven Reinforcement Learning via Variational Inference

Tim G. J. Rudner, Vitchyr Pong, Rowan McAllister, Yarin Gal, Sergey Levine

While reinforcement learning algorithms provide automated acquisition of optimal policies, practical application of such methods requires a number of design dec isions, such as manually designing reward functions that not only define the tas k, but also provide sufficient shaping to accomplish it. In this paper, we view reinforcement learning as inferring policies that achieve desired outcomes, rath er than as a problem of maximizing rewards. To solve this inference problem, we establish a novel variational inference formulation that allows us to derive a w ell-shaped reward function which can be learned directly from environment intera ctions. From the corresponding variational objective, we also derive a new proba bilistic Bellman backup operator and use it to develop an off-policy algorithm t o solve goal-directed tasks. We empirically demonstrate that this method elimina tes the need to hand-craft reward functions for a suite of diverse manipulation and locomotion tasks and leads to effective goal-directed behaviors.
****

Drawing Robust Scratch Tickets: Subnetworks with Inborn Robustness Are Found wit hin Randomly Initialized Networks

Yonggan Fu, Qixuan Yu, Yang Zhang, Shang Wu, Xu Ouyang, David Cox, Yingyan Lin

Deep Neural Networks (DNNs) are known to be vulnerable to adversarial attacks, i .e., an imperceptible perturbation to the input can mislead DNNs trained on clea n images into making erroneous predictions. To tackle this, adversarial training is currently the most effective defense method, by augmenting the training set with adversarial samples generated on the fly. \textbf{Interestingly, we discove r for the first time that there exist subnetworks with inborn robustness, matchi ng or surpassing the robust accuracy of the adversarially trained networks with comparable model sizes, within randomly initialized networks without any model t raining}, indicating that adversarial training on model weights is not indispens able towards adversarial robustness. We name such subnetworks Robust Scratch Tic kets (RSTs), which are also by nature efficient. Distinct from the popular lotte ry ticket hypothesis, neither the original dense networks nor the identified RST s need to be trained. To validate and understand this fascinating finding, we fu rther conduct extensive experiments to study the existence and properties of RST s under different models, datasets, sparsity patterns, and attacks, drawing insi ghts regarding the relationship between DNNs’ robustness and their initializatio n/overparameterization. Furthermore, we identify the poor adversarial transferab ility between RSTs of different sparsity ratios drawn from the same randomly ini tialized dense network, and propose a Random RST Switch (R2S) technique, which r andomly switches between different RSTs, as a novel defense method built on top of RSTs. We believe our findings about RSTs have opened up a new perspective to study model robustness and extend the lottery ticket hypothesis.
****

Rectifying the Shortcut Learning of Background for Few-Shot Learning

Xu Luo, Longhui Wei, Liangjian Wen, Jinrong Yang, Lingxi Xie, Zenglin Xu, Qi Tia n

The category gap between training and evaluation has been characterised as one o f the main obstacles to the success of Few-Shot Learning (FSL). In this paper, w e for the first time empirically identify image background, common in realistic

images, as a shortcut knowledge helpful for in-class classification but ungenera lizable beyond training categories in FSL. A novel framework, COSOC, is designed to tackle this problem by extracting foreground objects in images at both train ing and evaluation without any extra supervision. Extensive experiments carried on inductive FSL tasks demonstrate the effectiveness of our approaches.
****
SEAL: Self-supervised Embodied Active Learning using Exploration and 3D Consiste ncy
Devendra Singh Chaplot, Murtaza Dalal, Saurabh Gupta, Jitendra Malik, Russ R. Sa lakhutdinov
In this paper, we explore how we can build upon the data and models of Internet images and use them to adapt to robot vision without requiring any extra labels. We present a framework called Self-supervised Embodied Active Learning (SEAL). It utilizes perception models trained on internet images to learn an active expl oration policy. The observations gathered by this exploration policy are labelle d using 3D consistency and used to improve the perception model. We build and ut ilize 3D semantic maps to learn both action and perception in a completely self-supervised manner. The semantic map is used to compute an intrinsic motivation r eward for training the exploration policy and for labelling the agent observatio ns using spatio-temporal 3D consistency and label propagation. We demonstrate th at the SEAL framework can be used to close the action-perception loop: it improv es object detection and instance segmentation performance of a pretrained percep tion model by just moving around in training environments and the improved perce ption model can be used to improve Object Goal Navigation.
****
Sifting through the noise: Universal first-order methods for stochastic variatio nal inequalities
Kimon Antonakopoulos, Thomas Pethick, Ali Kavis, Panayotis Mertikopoulos, 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.
****
Accommodating Picky Customers: Regret Bound and Exploration Complexity for Multi -Objective Reinforcement Learning
Jingfeng Wu, Vladimir Braverman, 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.
****
Exact Privacy Guarantees for Markov Chain Implementations of the Exponential Mec hanism with Artificial Atoms
Jeremy Seeman, Matthew Reimherr, Aleksandra Slavkovi■
Requests for name changes in the electronic proceedings will be accepted with 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 Emergence of Objectness: Learning Zero-shot Segmentation from Videos
Runtao Liu, Zhirong Wu, Stella Yu, Stephen Lin
Humans can easily detect and segment moving objects simply by observing how they move, even without knowledge of object semantics. Inspired by this, we develop a zero-shot unsupervised approach for learning object segmentations. The model c omprises two visual pathways: an appearance pathway that segments individual RGB images into coherent object regions, and a motion pathway that predicts the flo w vector for each region between consecutive video frames. The two pathways join tly reconstruct a new representation called segment flow. This decoupled represe ntation of appearance and motion is trained in a self-supervised manner to recon

struct one frame from another.When pretrained on an unlabeled video corpus, the model can be useful for a variety of applications, including 1) primary object segmentation from a single image in a zero-shot fashion; 2) moving object segmentation from a video with unsupervised test-time adaptation; 3) image semantic segmentation by supervised fine-tuning on a labeled image dataset. We demonstrate encouraging experimental results on all of these tasks using pretrained models.
****

Direct Multi-view Multi-person 3D Pose Estimation
tao wang, Jianfeng Zhang, Yujun Cai, Shuicheng Yan, Jiashi Feng
We present Multi-view Pose transformer (MvP) for estimating multi-person 3D poses from multi-view images. Instead of estimating 3D joint locations from costly volumetric representation or reconstructing the per-person 3D pose from multiple detected 2D poses as in previous methods, MvP directly regresses the multi-person 3D poses in a clean and efficient way, without relying on intermediate tasks. Specifically, MvP represents skeleton joints as learnable query embeddings and let them progressively attend to and reason over the multi-view information from the input images to directly regress the actual 3D joint locations. To improve the accuracy of such a simple pipeline, MvP presents a hierarchical scheme to concisely represent query embeddings of multi-person skeleton joints and introduces an input-dependent query adaptation approach. Further, MvP designs a novel geometrically guided attention mechanism, called projective attention, to more precisely fuse the cross-view information for each joint. MvP also introduces a RayConv operation to integrate the view-dependent camera geometry into the feature representations for augmenting the projective attention. We show experimentally that our MvP model outperforms the state-of-the-art methods on several benchmarks while being much more efficient. Notably, it achieves 92.3% AP25 on the challenging Panoptic dataset, improving upon the previous best approach [35] by 9.8%. MvP is general and also extendable to recovering human mesh represented by the SMPL model, thus useful for modeling multi-person body shapes. Code and models are available at https://github.com/sail-sg/mvp.
****

MST: Masked Self-Supervised Transformer for Visual Representation
Zhaowen Li, Zhiyang Chen, Fan Yang, Wei Li, Yousong Zhu, Chaoyang Zhao, Rui Deng, Liwei Wu, Rui Zhao, Ming Tang, Jinqiao Wang
Transformer has been widely used for self-supervised pre-training in Natural Language Processing (NLP) and achieved great success. However, it has not been fully explored in visual self-supervised learning. Meanwhile, previous methods only consider the high-level feature and learning representation from a global perspective, which may fail to transfer to the downstream dense prediction tasks focusing on local features. In this paper, we present a novel Masked Self-supervised Transformer approach named MST, which can explicitly capture the local context of an image while preserving the global semantic information. Specifically, inspired by the Masked Language Modeling (MLM) in NLP, we propose a masked token strategy based on the multi-head self-attention map, which dynamically masks some tokens of local patches without damaging the crucial structure for self-supervised learning. More importantly, the masked tokens together with the remaining tokens are further recovered by a global image decoder, which preserves the spatial information of the image and is more friendly to the downstream dense prediction tasks. The experiments on multiple datasets demonstrate the effectiveness and generality of the proposed method. For instance, MST achieves Top-1 accuracy of 76.9% with DeiT-S only using 300-epoch pre-training by linear evaluation, which outperforms supervised methods with the same epoch by 0.4% and its comparable variant DINO by 1.0%. For dense prediction tasks, MST also achieves 42.7% mAP on MS COCO object detection and 74.04% mIoU on Cityscapes segmentation only with 100-epoch pre-training.
****

Exploiting Opponents Under Utility Constraints in Sequential Games
Martino Bernasconi-de-Luca, Federico Cacciamani, Simone Fioravanti, Nicola Gatti, Alberto Marchesi, Francesco Trovò
Recently, game-playing agents based on AI techniques have demonstrated super-hum

an performance in several sequential games, such as chess, Go, and poker. Surpri singly, the multi-agent learning techniques that allowed to reach these achievem ents do not take into account the actual behavior of the human player, potential ly leading to an impressive gap in performances. In this paper, we address the p roblem of designing artificial agents that learn how to effectively exploit unkn own human opponents while playing repeatedly against them in an online fashion. We study the case in which the agent's strategy during each repetition of the ga me is subject to constraints ensuring that the human's expected utility is withi n some lower and upper thresholds. Our framework encompasses several real-world problems, such as human engagement in repeated game playing and human education by means of serious games. As a first result, we formalize a set of linear inequ alities encoding the conditions that the agent's strategy must satisfy at each i teration in order to do not violate the given bounds for the human's expected ut ility. Then, we use such formulation in an upper confidence bound algorithm, and we prove that the resulting procedure suffers from sublinear regret and guarant ees that the constraints are satisfied with high probability at each iteration. Finally, we empirically evaluate the convergence of our algorithm on standard te stbeds of sequential games.
****
A Compositional Atlas of Tractable Circuit Operations for Probabilistic Inferenc e
Antonio Vergari, YooJung Choi, Anji Liu, Stefano Teso, Guy Van den Broeck
Circuit representations are becoming the lingua franca to express and reason abo ut tractable generative and discriminative models. In this paper, we show how co mplex inference scenarios for these models that commonly arise in machine learni ng---from computing the expectations of decision tree ensembles to information-theoretic divergences of sum-product networks---can be represented in terms of t ractable modular operations over circuits. Specifically, we characterize the tra ctability of simple transformations---sums, products, quotients, powers, logarit hms, and exponentials---in terms of sufficient structural constraints of the cir cuits they operate on, and present novel hardness results for the cases in which these properties are not satisfied. Building on these operations, we derive a u nified framework for reasoning about tractable models that generalizes several r esults in the literature and opens up novel tractable inference scenarios.
****
Demystifying and Generalizing BinaryConnect
Tim Dockhorn, Yaoliang Yu, Eyyüb Sari, Mahdi Zolnouri, Vahid Partovi Nia
BinaryConnect (BC) and its many variations have become the de facto standard for neural network quantization. However, our understanding of the inner workings o f BC is still quite limited. We attempt to close this gap in four different aspe cts: (a) we show that existing quantization algorithms, including post-training quantization, are surprisingly similar to each other; (b) we argue for proximal maps as a natural family of quantizers that is both easy to design and analyze; (c) we refine the observation that BC is a special case of dual averaging, which itself is a special case of the generalized conditional gradient algorithm; (d) consequently, we propose ProxConnect (PC) as a generalization of BC and we prov e its convergence properties by exploiting the established connections. We condu ct experiments on CIFAR-10 and ImageNet, and verify that PC achieves competitive performance.
****
CARMS: Categorical-Antithetic-REINFORCE Multi-Sample Gradient Estimator
Alek Dimitriev, Mingyuan Zhou
Accurately backpropagating the gradient through categorical variables is a chall enging task that arises in various domains, such as training discrete latent var iable models. To this end, we propose CARMS, an unbiased estimator for categoric al random variables based on multiple mutually negatively correlated (jointly an tithetic) samples. CARMS combines REINFORCE with copula based sampling to avoid duplicate samples and reduce its variance, while keeping the estimator unbiased using importance sampling. It generalizes both the ARMS antithetic estimator for binary variables, which is CARMS for two categories, as well as LOORF/VarGrad,

the leave-one-out REINFORCE estimator, which is CARMS with independent samples. We evaluate CARMS on several benchmark datasets on a generative modeling task, as well as a structured output prediction task, and find it to outperform competing methods including a strong self-control baseline. The code is publicly available.
****
Learning to Learn Dense Gaussian Processes for Few-Shot Learning
Ze Wang, Zichen Miao, Xiantong Zhen, Qiang Qiu
Gaussian processes with deep neural networks demonstrate to be a strong learner for few-shot learning since they combine the strength of deep learning and kernels while being able to well capture uncertainty. However, it remains an open problem to leverage the shared knowledge provided by related tasks. In this paper, we propose to learn Gaussian processes with dense inducing variables by meta-learning for few-shot learning. In contrast to sparse Gaussian processes, we define a set of dense inducing variables to be of a much larger size than the support set in each task, which collects prior knowledge from experienced tasks. The dense inducing variables specify a shared Gaussian process prior over prediction functions of all tasks, which are learned in a variational inference framework and offer a strong inductive bias for learning new tasks. To achieve task-specific prediction functions, we propose to adapt the inducing variables to each task by efficient gradient descent. We conduct extensive experiments on common benchmark datasets for a variety of few-shot learning tasks. Our dense Gaussian processes present significant improvements over vanilla Gaussian processes and comparable or even better performance with state-of-the-art methods.
****
Stochastic Solutions for Linear Inverse Problems using the Prior Implicit in a Denoiser
Zahra Kadkhodaie, Eero Simoncelli
Deep neural networks have provided state-of-the-art solutions for problems such as image denoising, which implicitly rely on a prior probability model of natural images. Two recent lines of work – Denoising Score Matching and Plug-and-Play – propose methodologies for drawing samples from this implicit prior and using it to solve inverse problems, respectively. Here, we develop a parsimonious and robust generalization of these ideas. We rely on a classic statistical result that shows the least-squares solution for removing additive Gaussian noise can be written directly in terms of the gradient of the log of the noisy signal density. We use this to derive a stochastic coarse-to-fine gradient ascent procedure for drawing high-probability samples from the implicit prior embedded within a CNN trained to perform blind denoising. A generalization of this algorithm to constrained sampling provides a method for using the implicit prior to solve any deterministic linear inverse problem, with no additional training, thus extending the power of supervised learning for denoising to a much broader set of problems. The algorithm relies on minimal assumptions and exhibits robust convergence over a wide range of parameter choices. To demonstrate the generality of our method, we use it to obtain state-of-the-art levels of unsupervised performance for deblurring, super-resolution, and compressive sensing.
****
Towards Stable and Robust AdderNets
Minjing Dong, Yunhe Wang, Xinghao Chen, Chang Xu
Adder neural network (AdderNet) replaces the original convolutions with massive multiplications by cheap additions while achieving comparable performance thus yields a series of energy-efficient neural networks. Compared with convolutional neural networks (CNNs), the training of AdderNets is much more sophisticated including several techniques for adjusting gradient and batch normalization. In addition, variances of both weights and activations in resulting adder networks are very enormous which limits its performance and the potential for applying to other tasks. To enhance the stability and robustness of AdderNets, we first thoroughly analyze the variance estimation of weight parameters and output features of an arbitrary adder layer. Then, we develop a weight normalization scheme for adaptively optimizing the weight distribution of AdderNets during the training pro

cedure, which can reduce the perturbation on running mean and variance in batch normalization layers. Meanwhile, the proposed weight normalization can also be utilized to enhance the adversarial robustness of resulting networks. Experiments conducted on several benchmarks demonstrate the superiority of the proposed approach for generating AdderNets with higher performance.
****
Representing Long-Range Context for Graph Neural Networks with Global Attention
Zhanghao Wu, Paras Jain, Matthew Wright, Azalia Mirhoseini, Joseph E. Gonzalez, Ion Stoica
Graph neural networks are powerful architectures for structured datasets. However, current methods struggle to represent long-range dependencies. Scaling the depth or width of GNNs is insufficient to broaden receptive fields as larger GNNs encounter optimization instabilities such as vanishing gradients and representation oversmoothing, while pooling-based approaches have yet to become as universally useful as in computer vision. In this work, we propose the use of Transformer-based self-attention to learn long-range pairwise relationships, with a novel “readout” mechanism to obtain a global graph embedding. Inspired by recent computer vision results that find position-invariant attention performant in learning long-range relationships, our method, which we call GraphTrans, applies a permutation-invariant Transformer module after a standard GNN module. This simple architecture leads to state-of-the-art results on several graph classification tasks, outperforming methods that explicitly encode graph structure. Our results suggest that purely-learning-based approaches without graph structure may be suitable for learning high-level, long-range relationships on graphs. Code for GraphTrans is available at https://github.com/ucbrise/graphtrans.
****
Beyond Bandit Feedback in Online Multiclass Classification
Dirk van der Hoeven, Federico Fusco, Nicolò Cesa-Bianchi
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 Student-Friendly Teacher Networks for Knowledge Distillation
Dae Young Park, Moon-Hyun Cha, changwook jeong, Daesin Kim, Bohyung Han
We propose a novel knowledge distillation approach to facilitate the transfer of dark knowledge from a teacher to a student. Contrary to most of the existing methods that rely on effective training of student models given pretrained teachers, we aim to learn the teacher models that are friendly to students and, consequently, more appropriate for knowledge transfer. In other words, at the time of optimizing a teacher model, the proposed algorithm learns the student branches jointly to obtain student-friendly representations. Since the main goal of our approach lies in training teacher models and the subsequent knowledge distillation procedure is straightforward, most of the existing knowledge distillation methods can adopt this technique to improve the performance of diverse student models in terms of accuracy and convergence speed. The proposed algorithm demonstrates outstanding accuracy in several well-known knowledge distillation techniques with various combinations of teacher and student models even in the case that their architectures are heterogeneous and there is no prior knowledge about student models at the time of training teacher networks
****
Implicit Transformer Network for Screen Content Image Continuous Super-Resolution
Jingyu Yang, Sheng Shen, Huanjing Yue, Kun Li
Nowadays, there is an explosive growth of screen contents due to the wide application of screen sharing, remote cooperation, and online education. To match the limited terminal bandwidth, high-resolution (HR) screen contents may be downsampled and compressed. At the receiver side, the super-resolution (SR)of low-resolution (LR) screen content images (SCIs) is highly demanded by the HR display or by the users to zoom in for detail observation. However, image SR methods most

ly designed for natural images do not generalize well for SCIs due to the very different image characteristics as well as the requirement of SCI browsing at arbitrary scales. To this end, we propose a novel Implicit Transformer Super-Resolution Network (ITSRN) for SCISR. For high-quality continuous SR at arbitrary ratios, pixel values at query coordinates are inferred from image features at key coordinates by the proposed implicit transformer and an implicit position encoding scheme is proposed to aggregate similar neighboring pixel values to the query one. We construct benchmark SCI1K and SCI1K-compression datasets withLR and HR SCI pairs. Extensive experiments show that the proposed ITSRN significantly outperforms several competitive continuous and discrete SR methods for both compressed and uncompressed SCIs.
****
Channel Permutations for N:M Sparsity
Jeff Pool, Chong Yu
We introduce channel permutations as a method to maximize the accuracy of N:M sparse networks. N:M sparsity requires N out of M consecutive elements to be zero and has been shown to maintain accuracy for many models and tasks with a simple prune and fine-tune workflow. By permuting weight matrices along their channel dimension and adjusting the surrounding layers appropriately, we demonstrate accuracy recovery for even small, parameter-efficient networks, without affecting inference run-time. We also present both a quality metric to simplify judging permutations as well as efficient methods to search for high-quality permutations, including two optimizations to escape local minima. Finally, we share an ablation study to show the importance of each part of our search algorithm, experimental results showing correlation between our quality metric and final network accuracy, improved sparse network accuracy using our techniques with insignificant overhead to training time, and the transformation of unstructured to structured sparse workloads. Code to use these techniques when generating a 2:4 sparse network is available at https://github.com/NVIDIA/apex/tree/master/apex/contrib/sparsity.
****
Curriculum Learning for Vision-and-Language Navigation
Jiwen Zhang, zhongyu wei, Jianqing Fan, Jiajie Peng
Vision-and-Language Navigation (VLN) is a task where an agent navigates in an embodied indoor environment under human instructions. Previous works ignore the distribution of sample difficulty and we argue that this potentially degrade their agent performance. To tackle this issue, we propose a novel curriculum- based training paradigm for VLN tasks that can balance human prior knowledge and agent learning progress about training samples. We develop the principle of curriculum design and re-arrange the benchmark Room-to-Room (R2R) dataset to make it suitable for curriculum training. Experiments show that our method is model-agnostic and can significantly improve the performance, the generalizability, and the training efficiency of current state-of-the-art navigation agents without increasing model complexity.
****
Better Algorithms for Individually Fair k-Clustering
Maryam Negahbani, Deeparnab Chakrabarty
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Video Instance Segmentation using Inter-Frame Communication Transformers
Sukjun Hwang, Miran Heo, Seoung Wug Oh, Seon Joo Kim
We propose a novel end-to-end solution for video instance segmentation (VIS) based on transformers. Recently, the per-clip pipeline shows superior performance over per-frame methods leveraging richer information from multiple frames. However, previous per-clip models require heavy computation and memory usage to achieve frame-to-frame communications, limiting practicality.In this work, we propose Inter-frame Communication Transformers (IFC), which significantly reduces the ov

erhead for information-passing between frames by efficiently encoding the contex t within the input clip.Specifically, we propose to utilize concise memory token s as a means of conveying information as well as summarizing each frame scene.Th e features of each frame are enriched and correlated with other frames through e xchange of information between the precisely encoded memory tokens.We validate o ur method on the latest benchmark sets and achieved state-of-the-art performance (AP 42.6 on YouTube-VIS 2019 val set using the offline inference) while having a considerably fast runtime (89.4 FPS). Our method can also be applied to near-o nline inference for processing a video in real-time with only a small delay.The code is available at https://github.com/sukjunhwang/IFC
****
Progressive Coordinate Transforms for Monocular 3D Object Detection
Li Wang, Li Zhang, Yi Zhu, Zhi Zhang, Tong He, Mu Li, Xiangyang Xue
Recognizing and localizing objects in the 3D space is a crucial ability for an A I agent to perceive its surrounding environment. While significant progress has been achieved with expensive LiDAR point clouds, it poses a great challenge for 3D object detection given only a monocular image. While there exist different al ternatives for tackling this problem, it is found that they are either equipped with heavy networks to fuse RGB and depth information or empirically ineffective to process millions of pseudo-LiDAR points. With in-depth examination, we reali ze that these limitations are rooted in inaccurate object localization. In this paper, we propose a novel and lightweight approach, dubbed {\em Progressive Coor dinate Transforms} (PCT) to facilitate learning coordinate representations. Spec ifically, a localization boosting mechanism with confidence-aware loss is introd uced to progressively refine the localization prediction. In addition, semantic image representation is also exploited to compensate for the usage of patch prop osals. Despite being lightweight and simple, our strategy allows us to establish a new state-of-the-art among the monocular 3D detectors on the competitive KITT I benchmark. At the same time, our proposed PCT shows great generalization to mo st coordinate-based 3D detection frameworks.
****
Structured Reordering for Modeling Latent Alignments in Sequence Transduction
bailin wang, Mirella Lapata, Ivan Titov
Despite success in many domains, neural models struggle in settings where train and test examples are drawn from different distributions. In particular, in cont rast to humans, conventional sequence-to-sequence (seq2seq) models fail to gener alize systematically, i.e., interpret sentences representing novel combinations of concepts (e.g., text segments) seen in training. Traditional grammar formalis ms excel in such settings by implicitly encoding alignments between input and ou tput segments, but are hard to scale and maintain. Instead of engineering a gra mmar, we directly model segment-to-segment alignments as discrete structured lat ent variables within a neural seq2seq model. To efficiently explore the large sp ace of alignments, we introduce a reorder-first align-later framework whose cent ral component is a neural reordering module producing separable permutations. We present an efficient dynamic programming algorithm performing exact marginal in ference of separable permutations, and, thus, enabling end-to-end differentiable training of our model. The resulting seq2seq model exhibits better systematic generalization than standard models on synthetic problems and NLP tasks (i.e., s emantic parsing and machine translation).
****
A universal probabilistic spike count model reveals ongoing modulation of neural variability
David Liu, Mate Lengyel
Neural responses are variable: even under identical experimental conditions, sin gle neuron and population responses typically differ from trial to trial and acr oss time. Recent work has demonstrated that this variability has predictable str ucture, can be modulated by sensory input and behaviour, and bears critical sign atures of the underlying network dynamics and computations. However, current met hods for characterising neural variability are primarily geared towards sensory coding in the laboratory: they require trials with repeatable experimental stimu

li and behavioural covariates. In addition, they make strong assumptions about the parametric form of variability, rely on assumption-free but data-inefficient histogram-based approaches, or are altogether ill-suited for capturing variability modulation by covariates. Here we present a universal probabilistic spike count model that eliminates these shortcomings. Our method builds on sparse Gaussian processes and can model arbitrary spike count distributions (SCDs) with flexible dependence on observed as well as latent covariates, using scalable variational inference to jointly infer the covariate-to-SCD mappings and latent trajectories in a data efficient way. Without requiring repeatable trials, it can flexibly capture covariate-dependent joint SCDs, and provide interpretable latent causes underlying the statistical dependencies between neurons. We apply the model to recordings from a canonical non-sensory neural population: head direction cells in the mouse. We find that variability in these cells defies a simple parametric relationship with mean spike count as assumed in standard models, its modulation by external covariates can be comparably strong to that of the mean firing rate, and slow low-dimensional latent factors explain away neural correlations. Our approach paves the way to understanding the mechanisms and computations underlying neural variability under naturalistic conditions, beyond the realm of sensory coding with repeatable stimuli.
****

Bellman Eluder Dimension: New Rich Classes of RL Problems, and Sample-Efficient Algorithms

Chi Jin, Qinghua Liu, Sobhan Miryoosefi

Finding the minimal structural assumptions that empower sample-efficient learning is one of the most important research directions in Reinforcement Learning (RL). This paper advances our understanding of this fundamental question by introducing a new complexity measure—Bellman Eluder (BE) dimension. We show that the family of RL problems of low BE dimension is remarkably rich, which subsumes a vast majority of existing tractable RL problems including but not limited to tabular MDPs, linear MDPs, reactive POMDPs, low Bellman rank problems as well as low Eluder dimension problems. This paper further designs a new optimization-based algorithm— GOLF, and reanalyzes a hypothesis elimination-based algorithm—OLIVE (proposed in Jiang et al. (2017)). We prove that both algorithms learn the near-optimal policies of low BE dimension problems in a number of samples that is polynomial in all relevant parameters, but independent of the size of state-action space. Our regret and sample complexity results match or improve the best existing results for several well-known subclasses of low BE dimension problems.
****

Detecting Anomalous Event Sequences with Temporal Point Processes

Oleksandr Shchur, Ali Caner Turkmen, Tim Januschowski, Jan Gasthaus, Stephan Günnemann

Automatically detecting anomalies in event data can provide substantial value in domains such as healthcare, DevOps, and information security. In this paper, we frame the problem of detecting anomalous continuous-time event sequences as out-of-distribution (OOD) detection for temporal point processes (TPPs). First, we show how this problem can be approached using goodness-of-fit (GoF) tests. We then demonstrate the limitations of popular GoF statistics for TPPs and propose a new test that addresses these shortcomings. The proposed method can be combined with various TPP models, such as neural TPPs, and is easy to implement. In our experiments, we show that the proposed statistic excels at both traditional GoF testing, as well as at detecting anomalies in simulated and real-world data.
****

HNPE: Leveraging Global Parameters for Neural Posterior Estimation

Pedro Rodrigues, Thomas Moreau, Gilles Louppe, Alexandre Gramfort

Inferring the parameters of a stochastic model based on experimental observations is central to the scientific method. A particularly challenging setting is when the model is strongly indeterminate, i.e. when distinct sets of parameters yield identical observations. This arises in many practical situations, such as when inferring the distance and power of a radio source (is the source close and weak or far and strong?) or when estimating the amplifier gain and underlying brai

n activity of an electrophysiological experiment. In this work, we present hiera
rchical neural posterior estimation (HNPE), a novel method for cracking such ind
eterminacy by exploiting additional information conveyed by an auxiliary set of
observations sharing global parameters. Our method extends recent developments i
n simulation-based inference (SBI) based on normalizing flows to Bayesian hierar
chical models. We validate quantitatively our proposal on a motivating example a
menable to analytical solutions and then apply it to invert a well known non-lin
ear model from computational neuroscience, using both simulated and real EEG dat
a.
****
Alignment Attention by Matching Key and Query Distributions
Shujian Zhang, Xinjie Fan, Huangjie Zheng, Korawat Tanwisuth, Mingyuan Zhou
The neural attention mechanism has been incorporated into deep neural networks t
o achieve state-of-the-art performance in various domains. Most such models use
multi-head self-attention which is appealing for the ability to attend to inform
ation from different perspectives. This paper introduces alignment attention tha
t explicitly encourages self-attention to match the distributions of the key and
 query within each head. The resulting alignment attention networks can be optim
ized as an unsupervised regularization in the existing attention framework. It i
s simple to convert any models with self-attention, including pre-trained ones,
to the proposed alignment attention. On a variety of language understanding task
s, we show the effectiveness of our method in accuracy, uncertainty estimation,
generalization across domains, and robustness to adversarial attacks. We further
 demonstrate the general applicability of our approach on graph attention and vi
sual question answering, showing the great potential of incorporating our alignm
ent method into various attention-related tasks.
****
Settling the Variance of Multi-Agent Policy Gradients
Jakub Grudzien Kuba, Muning Wen, Linghui Meng, shangding gu, Haifeng Zhang, Davi
d Mguni, Jun Wang, Yaodong Yang
Policy gradient (PG) methods are popular reinforcement learning (RL) methods whe
re a baseline is often applied to reduce the variance of gradient estimates. In
multi-agent RL (MARL), although the PG theorem can be naturally extended, the ef
fectiveness of multi-agent PG (MAPG) methods degrades as the variance of gradie
nt estimates increases rapidly with the number of agents. In this paper, we off
er a rigorous analysis of MAPG methods by, firstly, quantifying the contribution
s of the number of agents and agents' explorations to the variance of MAPG estim
ators. Based on this analysis, we derive the optimal baseline (OB) that achieves
 the minimal variance. In comparison to the OB, we measure the excess variance o
f existing MARL algorithms such as vanilla MAPG and COMA. Considering using deep
 neural networks, we also propose a surrogate version of OB, which can be seaml
essly plugged into any existing PG methods in MARL.   On benchmarks of Multi-Age
nt MuJoCo and StarCraft challenges, our OB technique effectively stabilises trai
ning and improves the performance of multi-agent PPO and COMA algorithms by a s
ignificant margin. Code is released at \url{https://github.com/morning9393/Opt
imal-Baseline-for-Multi-agent-Policy-Gradients}.
****
For high-dimensional hierarchical models, consider exchangeability of effects ac
ross covariates instead of across datasets
Brian Trippe, Hilary Finucane, Tamara Broderick
Hierarchical Bayesian methods enable information sharing across regression probl
ems on multiple groups of data. While standard practice is to model regression p
arameters (effects) as (1) exchangeable across the groups and (2) correlated to
differing degrees across covariates, we show that this approach exhibits poor st
atistical performance when the number of covariates exceeds the number of groups
. For instance, in statistical genetics, we might regress dozens of traits (defi
ning groups) for thousands of individuals (responses) on up to millions of genet
ic variants (covariates). When an analyst has more covariates than groups, we ar
gue that it is often preferable to instead model effects as (1) exchangeable acr
oss covariates and (2) correlated to differing degrees across groups. To this en

d, we propose a hierarchical model expressing our alternative perspective. We devise an empirical Bayes estimator for learning the degree of correlation between groups. We develop theory that demonstrates that our method outperforms the classic approach when the number of covariates dominates the number of groups, and corroborate this result empirically on several high-dimensional multiple regression and classification problems.
****
Efficient Algorithms for Learning Depth-2 Neural Networks with General ReLU Activations
Pranjal Awasthi, Alex Tang, Aravindan Vijayaraghavan
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Controllable and Compositional Generation with Latent-Space Energy-Based Models
Weili Nie, Arash Vahdat, Anima Anandkumar
Controllable generation is one of the key requirements for successful adoption of deep generative models in real-world applications, but it still remains as a great challenge. In particular, the compositional ability to generate novel concept combinations is out of reach for most current models. In this work, we use energy-based models (EBMs) to handle compositional generation over a set of attributes. To make them scalable to high-resolution image generation, we introduce an EBM in the latent space of a pre-trained generative model such as StyleGAN. We propose a novel EBM formulation representing the joint distribution of data and attributes together, and we show how sampling from it is formulated as solving an ordinary differential equation (ODE). Given a pre-trained generator, all we need for controllable generation is to train an attribute classifier. Sampling with ODEs is done efficiently in the latent space and is robust to hyperparameters. Thus, our method is simple, fast to train, and efficient to sample. Experimental results show that our method outperforms the state-of-the-art in both conditional sampling and sequential editing. In compositional generation, our method excels at zero-shot generation of unseen attribute combinations. Also, by composing energy functions with logical operators, this work is the first to achieve such compositionality in generating photo-realistic images of resolution 1024x1024.
****
Reverse-Complement Equivariant Networks for DNA Sequences
Vincent Mallet, Jean-Philippe Vert
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Provably Efficient Reinforcement Learning with Linear Function Approximation under Adaptivity Constraints
Tianhao Wang, Dongruo Zhou, 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.
****
Nonsmooth Implicit Differentiation for Machine-Learning and Optimization
Jérôme Bolte, Tam Le, Edouard Pauwels, Tony Silveti-Falls
In view of training increasingly complex learning architectures, we establish a nonsmooth implicit function theorem with an operational calculus. Our result applies to most practical problems (i.e., definable problems) provided that a nonsmooth form of the classical invertibility condition is fulfilled. This approach allows for formal subdifferentiation: for instance, replacing derivatives by Clarke Jacobians in the usual differentiation formulas is fully justified for a wide class of nonsmooth problems. Moreover this calculus is entirely compatible with

algorithmic differentiation (e.g., backpropagation). We provide several applications such as training deep equilibrium networks, training neural nets with conic optimization layers, or hyperparameter-tuning for nonsmooth Lasso-type models. To show the sharpness of our assumptions, we present numerical experiments showcasing the extremely pathological gradient dynamics one can encounter when applying implicit algorithmic differentiation without any hypothesis.
****
Heuristic-Guided Reinforcement Learning
Ching-An Cheng, Andrey Kolobov, Adith Swaminathan
We provide a framework to accelerate reinforcement learning (RL) algorithms by heuristics that are constructed by domain knowledge or offline data. Tabula rasa RL algorithms require environment interactions or computation that scales with the horizon of the sequential decision-making task. Using our framework, we show how heuristic-guided RL induces a much shorter horizon sub-problem that provably solves the original task. Our framework can be viewed as a horizon-based regularization for controlling bias and variance in RL under a finite interaction budget. In theory, we characterize the properties of a good heuristic and the resulting impact on RL acceleration. In particular, we introduce the novel concept of an improvable heuristic that can allow any RL agent to conservatively extrapolate beyond its prior knowledge. In practice, we instantiate our framework to accelerate several state-of-the-art algorithms in simulated robotic control tasks and procedurally generated games. Our framework complements the rich literature on warm-starting RL using expert demonstrations or exploratory data-sets, and creates a unified channel to inject prior knowledge into RL.
****
Statistical Undecidability in Linear, Non-Gaussian Causal Models in the Presence of Latent Confounders
Konstantin Genin
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 novel notion of barycenter for probability distributions based on optimal weak mass transport
Elsa Cazelles, Felipe Tobar, Joaquin Fontbona
We introduce weak barycenters of a family of probability distributions, based on the recently developed notion of optimal weak transport of mass by Gozlan et al. (2017) and Backhoff-Veraguas et al. (2020). We provide a theoretical analysis of this object and discuss its interpretation in the light of convex ordering between probability measures. In particular, we show that, rather than averaging the input distributions in a geometric way (as the Wasserstein barycenter based on classic optimal transport does) weak barycenters extract common geometric information shared by all the input distributions, encoded as a latent random variable that underlies all of them. We also provide an iterative algorithm to compute a weak barycenter for a finite family of input distributions, and a stochastic algorithm that computes them for arbitrary populations of laws. The latter approach is particularly well suited for the streaming setting, i.e., when distributions are observed sequentially. The notion of weak barycenter and our approaches to compute it are illustrated on synthetic examples, validated on 2D real-world data and compared to standard Wasserstein barycenters.
****
Temporal-attentive Covariance Pooling Networks for Video Recognition
Zilin Gao, Qilong Wang, Bingbing Zhang, Qinghua Hu, Peihua Li
For video recognition task, a global representation summarizing the whole contents of the video snippets plays an important role for the final performance. However, existing video architectures usually generate it by using a simple, global average pooling (GAP) method, which has limited ability to capture complex dynamics of videos. For image recognition task, there exist evidences showing that covariance pooling has stronger representation ability than GAP. Unfortunately, su

ch plain covariance pooling used in image recognition is an orderless representa tive, which cannot model spatio-temporal structure inherent in videos. Therefore , this paper proposes a Temporal-attentive Covariance Pooling (TCP), inserted at the end of deep architectures, to produce powerful video representations. Speci fically, our TCP first develops a temporal attention module to adaptively calibr ate spatio-temporal features for the succeeding covariance pooling, approximativ ely producing attentive covariance representations. Then, a temporal covariance pooling performs temporal pooling of the attentive covariance representations to characterize both intra-frame correlations and inter-frame cross-correlations o f the calibrated features. As such, the proposed TCP can capture complex tempora l dynamics. Finally, a fast matrix power normalization is introduced to exploit geometry of covariance representations. Note that our TCP is model-agnostic and can be flexibly integrated into any video architectures, resulting in TCPNet for effective video recognition. The extensive experiments on six benchmarks (e.g., Kinetics, Something-Something V1 and Charades) using various video architecture s show our TCPNet is clearly superior to its counterparts, while having strong g eneralization ability. The source code is publicly available.
****
Revisiting Smoothed Online Learning
Lijun Zhang, Wei Jiang, Shiyin Lu, Tianbao Yang
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-auth ors prior to requesting a name change in the electronic proceedings.
****
Marginalised Gaussian Processes with Nested Sampling
Fergus Simpson, Vidhi Lalchand, Carl Edward Rasmussen
Gaussian Process models are a rich distribution over functions with inductive bi ases controlled by a kernel function. Learning occurs through optimisation of th e kernel hyperparameters using the marginal likelihood as the objective. This wo rk proposes nested sampling as a means of marginalising kernel hyperparameters, because it is a technique that is well-suited to exploring complex, multi-modal distributions. We benchmark against Hamiltonian Monte Carlo on time-series and two-dimensional regression tasks, finding that a principled approach to quantify ing hyperparameter uncertainty substantially improves the quality of prediction intervals.
****
Provable Benefits of Actor-Critic Methods for Offline Reinforcement Learning
Andrea Zanette, Martin J Wainwright, Emma Brunskill
Actor-critic methods are widely used in offline reinforcement learningpractice, but are not so well-understood theoretically. We propose a newoffline actor-crit ic algorithm that naturally incorporates the pessimism principle, leading to sev eral key advantages compared to the state of the art. The algorithm can operate when the Bellman evaluation operator is closed with respect to the action value function of the actor's policies; this is a more general setting than the low-ra nk MDP model. Despite the added generality, the procedure is computationally tra ctable as it involves the solution of a sequence of second-order programs.We pro ve an upper bound on the suboptimality gap of the policy returned by the procedu re that depends on the data coverage of any arbitrary, possibly data dependent c omparator policy.The achievable guarantee is complemented with a minimax lower b ound that is matching up to logarithmic factors.
****
Bayesian Bellman Operators
Mattie Fellows, Kristian Hartikainen, Shimon Whiteson
We introduce a novel perspective on Bayesian reinforcement learning (RL); wherea s existing approaches infer a posterior over the transition distribution or Q-fu nction, we characterise the uncertainty in the Bellman operator. Our Bayesian Be llman operator (BBO) framework is motivated by the insight that when bootstrappi ng is introduced, model-free approaches actually infer a posterior over Bellman operators, not value functions. In this paper, we use BBO to provide a rigorous

theoretical analysis of model-free Bayesian RL to better understand its relation
ship to established frequentist RL methodologies. We prove that Bayesian solutio
ns are consistent with frequentist RL solutions, even when approximate inference
is used, and derive conditions for which convergence properties hold. Empirica
lly, we demonstrate that algorithms derived from the BBO framework have sophisti
cated deep exploration properties that enable them to solve continuous control t
asks at which state-of-the-art regularised actor-critic algorithms fail catastro
phically.
****
Uncertainty Calibration for Ensemble-Based Debiasing Methods
Ruibin Xiong, Yimeng Chen, Liang Pang, Xueqi Cheng, Zhi-Ming Ma, Yanyan Lan
Ensemble-based debiasing methods have been shown effective in mitigating the rel
iance of classifiers on specific dataset bias, by exploiting the output of a bia
s-only model to adjust the learning target. In this paper, we focus on the bias-
only model in these ensemble-based methods, which plays an important role but ha
s not gained much attention in the existing literature. Theoretically, we prove
that the debiasing performance can be damaged by inaccurate uncertainty estimati
ons of the bias-only model. Empirically, we show that existing bias-only models
fall short in producing accurate uncertainty estimations. Motivated by these fin
dings, we propose to conduct calibration on the bias-only model, thus achieving
a three-stage ensemble-based debiasing framework, including bias modeling, model
calibrating, and debiasing. Experimental results on NLI and fact verification t
asks show that our proposed three-stage debiasing framework consistently outperf
orms the traditional two-stage one in out-of-distribution accuracy.
****
Provably Faster Algorithms for Bilevel Optimization
Junjie Yang, Kaiyi Ji, 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-auth
ors prior to requesting a name change in the electronic proceedings.
****
Neo-GNNs: Neighborhood Overlap-aware Graph Neural Networks for Link Prediction
Seongjun Yun, Seoyoon Kim, Junhyun Lee, Jaewoo Kang, Hyunwoo J. Kim
Graph Neural Networks (GNNs) have been widely applied to various fields for lear
ning over graph-structured data. They have shown significant improvements over t
raditional heuristic methods in various tasks such as node classification and gr
aph classification. However, since GNNs heavily rely on smoothed node features r
ather than graph structure, they often show poor performance than simple heurist
ic methods in link prediction where the structural information, e.g., overlapped
neighborhoods, degrees, and shortest paths, is crucial. To address this limitat
ion, we propose Neighborhood Overlap-aware Graph Neural Networks (Neo-GNNs) that
learn useful structural features from an adjacency matrix and estimate overlapp
ed neighborhoods for link prediction. Our Neo-GNNs generalize neighborhood overl
ap-based heuristic methods and handle overlapped multi-hop neighborhoods. Our ex
tensive experiments on Open Graph Benchmark datasets (OGB) demonstrate that Neo-
GNNs consistently achieve state-of-the-art performance in link prediction.
****
Self-Supervised Multi-Object Tracking with Cross-input Consistency
Favyen Bastani, Songtao He, Samuel Madden
In this paper, we propose a self-supervised learning procedure for training a ro
bust multi-object tracking (MOT) model given only unlabeled video. While several
self-supervisory learning signals have been proposed in prior work on single-ob
ject tracking, such as color propagation and cycle-consistency, these signals ar
e not effective for training RNN models, which are needed to achieve accurate MO
T: they yield degenerate models that, for instance, always match new detections
to tracks with the closest initial detections. We propose a novel self-superviso
ry signal that we call cross-input consistency: we construct two distinct inputs
for the same sequence of video, by hiding different information about the seque
nce in each input. We then compute tracks in that sequence by applying an RNN mo

del independently on each input, and train the model to produce consistent track
s across the two inputs. We evaluate our unsupervised method on MOT17 and KITTI
--- remarkably, we find that, despite training only on unlabeled video, our unsu
pervised approach outperforms four supervised methods published in the last 1--2
 years, including Tracktor++, FAMNet, GSM, and mmMOT.
****
Tree in Tree: from Decision Trees to Decision Graphs
Bingzhao Zhu, Mahsa Shoaran
Decision trees have been widely used as classifiers in many machine learning app
lications thanks to their lightweight and interpretable decision process. This p
aper introduces Tree in Tree decision graph (TnT), a framework that extends the
conventional decision tree to a more generic and powerful directed acyclic graph
. TnT constructs decision graphs by recursively growing decision trees inside th
e internal or leaf nodes instead of greedy training. The time complexity of TnT
is linear to the number of nodes in the graph, therefore it can construct decisi
on graphs on large datasets. Compared to decision trees, we show that TnT achiev
es better classification performance with reduced model size, both as a stand-al
one classifier and as a base-estimator in bagging/AdaBoost ensembles. Our propos
ed model is a novel, more efficient and accurate alternative to the widely-used
decision trees.
****
Test-time Collective Prediction
Celestine Mendler-Dünner, Wenshuo Guo, Stephen Bates, Michael Jordan
An increasingly common setting in machine learning involves multiple parties, ea
ch with their own data, who want to jointly make predictions on future test poin
ts. Agents wish to benefit from the collective expertise of the full set of agen
ts to make better predictions than they would individually, but may not be willi
ng to release labeled data or model parameters. In this work, we explore a decen
tralized mechanism to make collective predictions at test time, that is inspired
 by the literature in social science on human consensus-making. Building on a qu
ery model to facilitate information exchange among agents, our approach leverage
s each agent’s pre-trained model without relying on external validation, model r
etraining, or data pooling. A theoretical analysis shows that our approach recov
ers inverse mean-squared-error (MSE) weighting in the large-sample limit which i
s known to be the optimal way to combine independent, unbiased estimators. Empir
ically, we demonstrate that our scheme effectively combines models with differin
g quality across the input space: the proposed consensus prediction achieves sig
nificant gains over classical model averaging, and even outperforms weighted ave
raging schemes that have access to additional validation data. Finally, we propo
se a decentralized Jackknife procedure as a tool to evaluate the sensitivity of
the collective predictions with respect to a single agent's opinion.
****
A Continuous Mapping For Augmentation Design
Keyu Tian, Chen Lin, Ser Nam Lim, Wanli Ouyang, Puneet Dokania, Philip Torr
Requests for name changes in the electronic proceedings will be accepted with 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 Routing by Memory
Kaipeng Zhang, Zhenqiang Li, Zhifeng Li, Wei Liu, Yoichi Sato
Recent Convolutional Neural Networks (CNNs) have achieved significant success by
 stacking multiple convolutional blocks, named procedures in this paper, to extr
act semantic features. However, they use the same procedure sequence for all inp
uts, regardless of the intermediate features.This paper proffers a simple yet ef
fective idea of constructing parallel procedures and assigning similar intermedi
ate features to the same specialized procedures in a divide-and-conquer fashion.
 It relieves each procedure's learning difficulty and thus leads to superior per
formance. Specifically, we propose a routing-by-memory mechanism for existing CN
N architectures. In each stage of the network, we introduce parallel Procedural

Units (PUs). A PU consists of a memory head and a procedure. The memory head mai
ntains a summary of a type of features. For an intermediate feature, we search i
ts closest memory and forward it to the corresponding procedure in both training
 and testing. In this way, different procedures are tailored to different featur
es and therefore tackle them better.Networks with the proposed mechanism can be
trained efficiently using a four-step training strategy. Experimental results sh
ow that our method improves VGGNet, ResNet, and EfficientNet's accuracies on Tin
y ImageNet, ImageNet, and CIFAR-100 benchmarks with a negligible extra computati
onal cost.
****
GeoMol: Torsional Geometric Generation of Molecular 3D Conformer Ensembles
Octavian Ganea, Lagnajit Pattanaik, Connor Coley, Regina Barzilay, Klavs Jensen,
 William Green, Tommi Jaakkola
Prediction of a molecule’s 3D conformer ensemble from the molecular graph holds
a key role in areas of cheminformatics and drug discovery. Existing generative m
odels have several drawbacks including lack of modeling important molecular geom
etry elements (e.g., torsion angles), separate optimization stages prone to erro
r accumulation, and the need for structure fine-tuning based on approximate clas
sical force-fields or computationally expensive methods. We propose GEOMOL --- a
n end-to-end, non-autoregressive, and SE(3)-invariant machine learning approach
to generate distributions of low-energy molecular 3D conformers. Leveraging the
power of message passing neural networks (MPNNs) to capture local and global gra
ph information, we predict local atomic 3D structures and torsion angles, avoid-
 ing unnecessary over-parameterization of the geometric degrees of freedom (e.g.
, one angle per non-terminal bond). Such local predictions suffice both for both
 the training loss computation and for the full deterministic conformer assembly
 (at test time). We devise a non-adversarial optimal transport based loss functi
on to promote diverse conformer generation. GEOMOL predominantly outperforms pop
ular open-source, commercial, or state-of-the-art machine learning (ML) models,
while achieving significant speed-ups. We expect such differentiable 3D structur
e generators to significantly impact molecular modeling and related applications
.
****
CANITA: Faster Rates for Distributed Convex Optimization with Communication Comp
ression
Zhize Li, 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.
****
Drop-DTW: Aligning Common Signal Between Sequences While Dropping Outliers
Mikita Dvornik, Isma Hadji, Konstantinos G. Derpanis, Animesh Garg, Allan Jepson
In this work, we consider the problem of sequence-to-sequence alignment for sign
als containing outliers. Assuming the absence of outliers, the standard Dynamic
 Time Warping (DTW) algorithm efficiently computes the optimal alignment between
 two (generally) variable-length sequences. While DTW is robust to temporal shi
fts and dilations of the signal, it fails to align sequences in a meaningful way
 in the presence of outliers that can be arbitrarily interspersed in the sequenc
es. To address this problem, we introduce Drop-DTW, a novel algorithm that alig
ns the common signal between the sequences while automatically dropping the outl
ier elements from the matching. The entire procedure is implemented as a single
 dynamic program that is efficient and fully differentiable. In our experiments
, we show that Drop-DTW is a robust similarity measure for sequence retrieval an
d demonstrate its effectiveness as a training loss on diverse applications. With
 Drop-DTW, we address temporal step localization on instructional videos, repres
entation learning from noisy videos, and cross-modal representation learning for
 audio-visual retrieval and localization. In all applications, we take a weakly-
 or unsupervised approach and demonstrate state-of-the-art results under these s
ettings.

****
Safe Reinforcement Learning with Natural Language Constraints
Tsung-Yen Yang, Michael Y Hu, Yinlam Chow, Peter J Ramadge, Karthik Narasimhan
While safe reinforcement learning (RL) holds great promise for many practical ap plications like robotics or autonomous cars, current approaches require specifyi ng constraints in mathematical form. Such specifications demand domain expertise , limiting the adoption of safe RL. In this paper, we propose learning to interp ret natural language constraints for safe RL. To this end, we first introduce HA ZARDWORLD, a new multi-task benchmark that requires an agent to optimize reward while not violating constraints specified in free-form text. We then develop an agent with a modular architecture that can interpret and adhere to such textual constraints while learning new tasks. Our model consists of (1) a constraint int erpreter that encodes textual constraints into spatial and temporal representati ons of forbidden states, and (2) a policy network that uses these representation s to produce a policy achieving minimal constraint violations during training. A cross different domains in HAZARDWORLD, we show that our method achieves higher rewards (up to11x) and fewer constraint violations (by 1.8x) compared to existin g approaches. However, in terms of absolute performance, HAZARDWORLD still poses significant challenges for agents to learn efficiently, motivating the need for future work.
****
Compositional Modeling of Nonlinear Dynamical Systems with ODE-based Random Feat ures
Thomas McDonald, Mauricio Álvarez
Effectively modeling phenomena present in highly nonlinear dynamical systems whi lst also accurately quantifying uncertainty is a challenging task, which often r equires problem-specific techniques. We present a novel, domain-agnostic approac h to tackling this problem, using compositions of physics-informed random featur es, derived from ordinary differential equations. The architecture of our model leverages recent advances in approximate inference for deep Gaussian processes, such as layer-wise weight-space approximations which allow us to incorporate ran dom Fourier features, and stochastic variational inference for approximate Bayes ian inference. We provide evidence that our model is capable of capturing highly nonlinear behaviour in real-world multivariate time series data. In addition, w e find that our approach achieves comparable performance to a number of other pr obabilistic models on benchmark regression tasks.
****
Implicit Semantic Response Alignment for Partial Domain Adaptation
Wenxiao Xiao, Zhengming Ding, Hongfu Liu
Partial Domain Adaptation (PDA) addresses the unsupervised domain adaptation pro blem where the target label space is a subset of the source label space. Most st ate-of-art PDA methods tackle the inconsistent label space by assigning weights to classes or individual samples, in an attempt to discard the source data that belongs to the irrelevant classes. However, we believe samples from those extra categories would still contain valuable information to promote positive transfer . In this paper, we propose the Implicit Semantic Response Alignment to explore the intrinsic relationships among different categories by applying a weighted sc hema on the feature level. Specifically, we design a class2vec module to extract the implicit semantic topics from the visual features. With an attention layer, we calculate the semantic response according to each implicit semantic topic. T hen semantic responses of source and target data are aligned to retain the relev ant information contained in multiple categories by weighting the features, inst ead of samples. Experiments on several cross-domain benchmark datasets demonstra te the effectiveness of our method over the state-of-the-art PDA methods. Moreov er, we elaborate in-depth analyses to further explore implicit semantic alignmen t.
****
ToAlign: Task-Oriented Alignment for Unsupervised Domain Adaptation
Guoqiang Wei, Cuiling Lan, Wenjun Zeng, Zhizheng Zhang, Zhibo Chen
Unsupervised domain adaptive classifcation intends to improve the classifcation

performance on unlabeled target domain. To alleviate the adverse effect of domai
n shift, many approaches align the source and target domains in the feature spac
e. However, a feature is usually taken as a whole for alignment without explicit
ly making domain alignment proactively serve the classifcation task, leading to
sub-optimal solution. In this paper, we propose an effective Task-oriented Align
ment (ToAlign) for unsupervised domain adaptation (UDA). We study what features
should be aligned across domains and propose to make the domain alignment proact
ively serve classifcation by performing feature decomposition and alignment unde
r the guidance of the prior knowledge induced from the classifcation task itself
. Particularly, we explicitly decompose a feature in the source domain into a ta
sk-related/discriminative feature that should be aligned, and a task-irrelevant
feature that should be avoided/ignored, based on the classifcation meta-knowledg
e. Extensive experimental results on various benchmarks (e.g., Offce-Home, Visda
-2017, and DomainNet) under different domain adaptation settings demonstrate the
 effectiveness of ToAlign which helps achieve the state-of-the-art performance.
The code is publicly available at https://github.com/microsoft/UDA.
****
Prior-independent Dynamic Auctions for a Value-maximizing Buyer
Yuan Deng, Hanrui Zhang
Requests for name changes in the electronic proceedings will be accepted with no
 questions asked. However name changes may cause bibliographic tracking issues.
  Authors are asked to consider this carefully and discuss it with their co-auth
ors prior to requesting a name change in the electronic proceedings.
****
Safe Reinforcement Learning by Imagining the Near Future
Garrett Thomas, Yuping Luo, Tengyu Ma
Safe reinforcement learning is a promising path toward applying reinforcement le
arning algorithms to real-world problems, where suboptimal behaviors may lead to
 actual negative consequences. In this work, we focus on the setting where unsaf
e states can be avoided by planning ahead a short time into the future. In this
setting, a model-based agent with a sufficiently accurate model can avoid unsafe
 states.We devise a model-based algorithm that heavily penalizes unsafe trajecto
ries, and derive guarantees that our algorithm can avoid unsafe states under cer
tain assumptions. Experiments demonstrate that our algorithm can achieve competi
tive rewards with fewer safety violations in several continuous control tasks.
****
Contrastive Active Inference
Pietro Mazzaglia, Tim Verbelen, Bart Dhoedt
Active inference is a unifying theory for perception and action resting upon the
 idea that the brain maintains an internal model of the world by minimizing free
 energy. From a behavioral perspective, active inference agents can be seen as s
elf-evidencing beings that act to fulfill their optimistic predictions, namely p
referred outcomes or goals. In contrast, reinforcement learning requires human-d
esigned rewards to accomplish any desired outcome. Although active inference cou
ld provide a more natural self-supervised objective for control, its applicabili
ty has been limited because of the shortcomings in scaling the approach to compl
ex environments. In this work, we propose a contrastive objective for active inf
erence that strongly reduces the computational burden in learning the agent's ge
nerative model and planning future actions. Our method performs notably better t
han likelihood-based active inference in image-based tasks, while also being com
putationally cheaper and easier to train. We compare to reinforcement learning a
gents that have access to human-designed reward functions, showing that our appr
oach closely matches their performance. Finally, we also show that contrastive m
ethods perform significantly better in the case of distractors in the environmen
t and that our method is able to generalize goals to variations in the backgroun
d.
****
Overparameterization Improves Robustness to Covariate Shift in High Dimensions
Nilesh Tripuraneni, Ben Adlam, Jeffrey Pennington
A significant obstacle in the development of robust machine learning models is \

emph{covariate shift}, a form of distribution shift that occurs when the input d istributions of the training and test sets differ while the conditional label di stributions remain the same. Despite the prevalence of covariate shift in real-w orld applications, a theoretical understanding in the context of modern machine learning has remained lacking. In this work, we examine the exact high-dimension al asymptotics of random feature regression under covariate shift and present a precise characterization of the limiting test error, bias, and variance in this setting. Our results motivate a natural partial order over covariate shifts that provides a sufficient condition for determining when the shift will harm (or ev en help) test performance. We find that overparameterized models exhibit enhance d robustness to covariate shift, providing one of the first theoretical explanat ions for this ubiquitous empirical phenomenon. Additionally, our analysis reveal s an exact linear relationship between the in-distribution and out-of-distributi on generalization performance, offering an explanation for this surprising recen t observation.
****
Logarithmic Regret in Feature-based Dynamic Pricing
Jianyu Xu, 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.
****
Dimension-free empirical entropy estimation
Doron Cohen, Aryeh Kontorovich, Aaron Koolyk, Geoffrey Wolfer
We seek an entropy estimator for discrete distributions with fully empirical acc uracy bounds. As stated, this goal is infeasible without some prior assumptions on the distribution. We discover that a certain information moment assumption re nders the problem feasible. We argue that the moment assumption is natural and, in some sense, {\em minimalistic} --- weaker than finite support or tail decay c onditions. Under the moment assumption, we provide the first finite-sample entro py estimates for infinite alphabets, nearly recovering the known minimax rates. Moreover, we demonstrate that our empirical bounds are significantly sharper tha n the state-of-the-art bounds, for various natural distributions and non-trivial sample regimes. Along the way, we give a dimension-free analogue of the Cover-T homas result on entropy continuity (with respect to total variation distance) fo r finite alphabets, which may be of independent interest.
****
Towards Biologically Plausible Convolutional Networks
Roman Pogodin, Yash Mehta, Timothy Lillicrap, Peter E Latham
Convolutional networks are ubiquitous in deep learning. They are particularly us eful for images, as they reduce the number of parameters, reduce training time, and increase accuracy. However, as a model of the brain they are seriously probl ematic, since they require weight sharing - something real neurons simply cannot do. Consequently, while neurons in the brain can be locally connected (one of t he features of convolutional networks), they cannot be convolutional. Locally co nnected but non-convolutional networks, however, significantly underperform conv olutional ones. This is troublesome for studies that use convolutional networks to explain activity in the visual system. Here we study plausible alternatives t o weight sharing that aim at the same regularization principle, which is to make each neuron within a pool react similarly to identical inputs. The most natural way to do that is by showing the network multiple translations of the same imag e, akin to saccades in animal vision. However, this approach requires many trans lations, and doesn't remove the performance gap. We propose instead to add later al connectivity to a locally connected network, and allow learning via Hebbian p lasticity. This requires the network to pause occasionally for a sleep-like phas e of "weight sharing". This method enables locally connected networks to achieve nearly convolutional performance on ImageNet and improves their fit to the vent ral stream data, thus supporting convolutional networks as a model of the visual stream.

****
DynamicViT: Efficient Vision Transformers with Dynamic Token Sparsification
Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, 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 Transferable Adversarial Perturbations
Krishna kanth Nakka, Mathieu Salzmann
While effective, deep neural networks (DNNs) are vulnerable to adversarial attacks. In particular, recent work has shown that such attacks could be generated by another deep network, leading to significant speedups over optimization-based perturbations. However, the ability of such generative methods to generalize to different test-time situations has not been systematically studied. In this paper, we, therefore, investigate the transferability of generated perturbations when the conditions at inference time differ from the training ones in terms of the target architecture, target data, and target task. Specifically, we identify the mid-level features extracted by the intermediate layers of DNNs as common ground across different architectures, datasets, and tasks. This lets us introduce a loss function based on such mid-level features to learn an effective, transferable perturbation generator. Our experiments demonstrate that our approach outperforms the state-of-the-art universal and transferable attack strategies.
****
PortaSpeech: Portable and High-Quality Generative Text-to-Speech
Yi Ren, Jinglin Liu, Zhou Zhao
Non-autoregressive text-to-speech (NAR-TTS) models such as FastSpeech 2 and Glow-TTS can synthesize high-quality speech from the given text in parallel. After analyzing two kinds of generative NAR-TTS models (VAE and normalizing flow), we find that: VAE is good at capturing the long-range semantics features (e.g., prosody) even with small model size but suffers from blurry and unnatural results; and normalizing flow is good at reconstructing the frequency bin-wise details but performs poorly when the number of model parameters is limited. Inspired by these observations, to generate diverse speech with natural details and rich prosody using a lightweight architecture, we propose PortaSpeech, a portable and high-quality generative text-to-speech model. Specifically, 1) to model both the prosody and mel-spectrogram details accurately, we adopt a lightweight VAE with an enhanced prior followed by a flow-based post-net with strong conditional inputs as the main architecture. 2) To further compress the model size and memory footprint, we introduce the grouped parameter sharing mechanism to the affine coupling layers in the post-net. 3) To improve the expressiveness of synthesized speech and reduce the dependency on accurate fine-grained alignment between text and speech, we propose a linguistic encoder with mixture alignment combining hard word-level alignment and soft phoneme-level alignment, which explicitly extracts word-level semantic information. Experimental results show that PortaSpeech outperforms other TTS models in both voice quality and prosody modeling in terms of subjective and objective evaluation metrics, and shows only a slight performance degradation when reducing the model parameters to 6.7M (about 4x model size and 3x runtime memory compression ratio compared with FastSpeech 2). Our extensive ablation studies demonstrate that each design in PortaSpeech is effective.
****
Exponential Graph is Provably Efficient for Decentralized Deep Training
Bicheng Ying, Kun Yuan, Yiming Chen, Hanbin Hu, PAN PAN, Wotao Yin
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
CLIP-It! Language-Guided Video Summarization
Medhini Narasimhan, Anna Rohrbach, Trevor Darrell

A generic video summary is an abridged version of a video that conveys the whole story and features the most important scenes. Yet the importance of scenes in a video is often subjective, and users should have the option of customizing the summary by using natural language to specify what is important to them. Further, existing models for fully automatic generic summarization have not exploited available language models, which can serve as an effective prior for saliency. This work introduces CLIP-It, a single framework for addressing both generic and query-focused video summarization, typically approached separately in the literature. We propose a language-guided multimodal transformer that learns to score frames in a video based on their importance relative to one another and their correlation with a user-defined query (for query-focused summarization) or an automatically generated dense video caption (for generic video summarization). Our model can be extended to the unsupervised setting by training without ground-truth supervision. We outperform baselines and prior work by a significant margin on both standard video summarization datasets (TVSum and SumMe) and a query-focused video summarization dataset (QFVS). Particularly, we achieve large improvements in the transfer setting, attesting to our method's strong generalization capabilities.
****
Learning Treatment Effects in Panels with General Intervention Patterns
Vivek Farias, Andrew Li, Tianyi 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-authors prior to requesting a name change in the electronic proceedings.
****
Lossy Compression for Lossless Prediction
Yann Dubois, Benjamin Bloem-Reddy, Karen Ullrich, Chris J. Maddison
Most data is automatically collected and only ever "seen" by algorithms. Yet, data compressors preserve perceptual fidelity rather than just the information needed by algorithms performing downstream tasks. In this paper, we characterize the bit-rate required to ensure high performance on all predictive tasks that are invariant under a set of transformations, such as data augmentations. Based on our theory, we design unsupervised objectives for training neural compressors. Using these objectives, we train a generic image compressor that achieves substantial rate savings (more than 1000x on ImageNet) compared to JPEG on 8 datasets, without decreasing downstream classification performance.
****
From Optimality to Robustness: Adaptive Re-Sampling Strategies in Stochastic Bandits
Dorian Baudry, Patrick Saux, Odalric-Ambrym Maillard
The stochastic multi-arm bandit problem has been extensively studied under standard assumptions on the arm's distribution (e.g bounded with known support, exponential family, etc). These assumptions are suitable for many real-world problems but sometimes they require knowledge (on tails for instance) that may not be precisely accessible to the practitioner, raising the question of the robustness of bandit algorithms to model misspecification. In this paper we study a generic \emph{Dirichlet Sampling} (DS) algorithm, based on pairwise comparisons of empirical indices computed with \textit{re-sampling} of the arms' observations and a data-dependent \textit{exploration bonus}. We show that different variants of this strategy achieve provably optimal regret guarantees when the distributions are bounded and logarithmic regret for semi-bounded distributions with a mild quantile condition. We also show that a simple tuning achieve robustness with respect to a large class of unbounded distributions, at the cost of slightly worse than logarithmic asymptotic regret. We finally provide numerical experiments showing the merits of DS in a decision-making problem on synthetic agriculture data.
****
CCVS: Context-aware Controllable Video Synthesis
Guillaume Le Moing, Jean Ponce, Cordelia Schmid
This presentation introduces a self-supervised learning approach to the synthesi

s of new videos clips from old ones, with several new key elements for improved spatial resolution and realism: It conditions the synthesis process on contextual information for temporal continuity and ancillary information for fine control. The prediction model is doubly autoregressive, in the latent space of an autoencoder for forecasting, and in image space for updating contextual information, which is also used to enforce spatio-temporal consistency through a learnable optical flow module. Adversarial training of the autoencoder in the appearance and temporal domains is used to further improve the realism of its output. A quantizer inserted between the encoder and the transformer in charge of forecasting future frames in latent space (and its inverse inserted between the transformer and the decoder) adds even more flexibility by affording simple mechanisms for handling multimodal ancillary information for controlling the synthesis process (e.g., a few sample frames, an audio track, a trajectory in image space) and taking into account the intrinsically uncertain nature of the future by allowing multiple predictions. Experiments with an implementation of the proposed approach give very good qualitative and quantitative results on multiple tasks and standard benchmarks.
****
An Online Riemannian PCA for Stochastic Canonical Correlation Analysis
Zihang Meng, Rudrasis Chakraborty, Vikas Singh
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Predify: Augmenting deep neural networks with brain-inspired predictive coding dynamics
Bhavin Choksi, Milad Mozafari, Callum Biggs O'May, B. ADOR, Andrea Alamia, Rufin VanRullen
Deep neural networks excel at image classification, but their performance is far less robust to input perturbations than human perception. In this work we explore whether this shortcoming may be partly addressed by incorporating brain-inspired recurrent dynamics in deep convolutional networks. We take inspiration from a popular framework in neuroscience: "predictive coding". At each layer of the hierarchical model, generative feedback "predicts" (i.e., reconstructs) the pattern of activity in the previous layer. The reconstruction errors are used to iteratively update the network’s representations across timesteps, and to optimize the network's feedback weights over the natural image dataset--a form of unsupervised training. We show that implementing this strategy into two popular networks, VGG16 and EfficientNetB0, improves their robustness against various corruptions and adversarial attacks. We hypothesize that other feedforward networks could similarly benefit from the proposed framework. To promote research in this direction, we provide an open-sourced PyTorch-based package called \textit{Predify}, which can be used to implement and investigate the impacts of the predictive coding dynamics in any convolutional neural network.
****
Deep Extrapolation for Attribute-Enhanced Generation
Alvin Chan, Ali Madani, Ben Krause, Nikhil Naik
Attribute extrapolation in sample generation is challenging for deep neural networks operating beyond the training distribution. We formulate a new task for extrapolation in sequence generation, focusing on natural language and proteins, and propose GENhance, a generative framework that enhances attributes through a learned latent space. Trained on movie reviews and a computed protein stability dataset, GENhance can generate strongly-positive text reviews and highly stable protein sequences without being exposed to similar data during training. We release our benchmark tasks and models to contribute to the study of generative modeling extrapolation and data-driven design in biology and chemistry.
****
Generalized DataWeighting via Class-Level Gradient Manipulation
Can Chen, Shuhao Zheng, Xi Chen, Erqun Dong, Xue (Steve) Liu, Hao Liu, Dejing Do

u
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Slow Learning and Fast Inference: Efficient Graph Similarity Computation via Knowledge Distillation
Can Qin, Handong Zhao, Lichen Wang, Huan Wang, Yulun Zhang, Yun Fu
Graph Similarity Computation (GSC) is essential to wide-ranging graph applications such as retrieval, plagiarism/anomaly detection, etc. The exact computation of graph similarity, e.g., Graph Edit Distance (GED), is an NP-hard problem that cannot be exactly solved within an adequate time given large graphs. Thanks to the strong representation power of graph neural network (GNN), a variety of GNN-based inexact methods emerged. To capture the subtle difference across graphs, the key success is designing the dense interaction with features fusion at the early stage, which, however, is a trade-off between speed and accuracy. For slow learning of graph similarity, this paper proposes a novel early-fusion approach by designing a co-attention-based feature fusion network on multilevel GNN features. To further improve the speed without much accuracy drop, we introduce an efficient GSC solution by distilling the knowledge from the slow early-fusion model to the student one for fast inference. Such a student model also enables the offline collection of individual graph embeddings, speeding up the inference time in orders. To address the instability through knowledge transfer, we decompose the dynamic joint embedding into the static pseudo individual ones for precise teacher-student alignment. The experimental analysis on the real-world datasets demonstrates the superiority of our approach over the state-of-the-art methods on both accuracy and efficiency. Particularly, we speed up the prior art by more than 10x on the benchmark AIDS data.
****
Meta Learning Backpropagation And Improving It
Louis Kirsch, Jürgen Schmidhuber
Many concepts have been proposed for meta learning with neural networks (NNs), e.g., NNs that learn to reprogram fast weights, Hebbian plasticity, learned learning rules, and meta recurrent NNs. Our Variable Shared Meta Learning (VSML) unifies the above and demonstrates that simple weight-sharing and sparsity in an NN is sufficient to express powerful learning algorithms (LAs) in a reusable fashion. A simple implementation of VSML where the weights of a neural network are replaced by tiny LSTMs allows for implementing the backpropagation LA solely by running in forward-mode. It can even meta learn new LAs that differ from online backpropagation and generalize to datasets outside of the meta training distribution without explicit gradient calculation. Introspection reveals that our meta learned LAs learn through fast association in a way that is qualitatively different from gradient descent.
****
Posterior Meta-Replay for Continual Learning
Christian Henning, Maria Cervera, Francesco D'Angelo, Johannes von Oswald, Regina Traber, Benjamin Ehret, Seijin Kobayashi, Benjamin F. Grewe, João Sacramento
Learning a sequence of tasks without access to i.i.d. observations is a widely studied form of continual learning (CL) that remains challenging. In principle, Bayesian learning directly applies to this setting, since recursive and one-off Bayesian updates yield the same result. In practice, however, recursive updating often leads to poor trade-off solutions across tasks because approximate inference is necessary for most models of interest. Here, we describe an alternative Bayesian approach where task-conditioned parameter distributions are continually inferred from data. We offer a practical deep learning implementation of our framework based on probabilistic task-conditioned hypernetworks, an approach we term posterior meta-replay. Experiments on standard benchmarks show that our probabilistic hypernetworks compress sequences of posterior parameter distributions with virtually no forgetting. We obtain considerable performance gains compared to

existing Bayesian CL methods, and identify task inference as our major limiting factor. This limitation has several causes that are independent of the considere d sequential setting, opening up new avenues for progress in CL.
****
Optimizing Reusable Knowledge for Continual Learning via Metalearning
Julio Hurtado, Alain Raymond, Alvaro Soto
When learning tasks over time, artificial neural networks suffer from a problem known as Catastrophic Forgetting (CF). This happens when the weights of a networ k are overwritten during the training of a new task causing forgetting of old in formation. To address this issue, we propose MetA Reusable Knowledge or MARK, a new method that fosters weight reusability instead of overwriting when learning a new task. Specifically, MARK keeps a set of shared weights among tasks. We env ision these shared weights as a common Knowledge Base (KB) that is not only used to learn new tasks, but also enriched with new knowledge as the model learns ne w tasks. Key components behind MARK are two-fold. On the one hand, a metalearnin g approach provides the key mechanism to incrementally enrich the KB with new kn owledge and to foster weight reusability among tasks. On the other hand, a set o f trainable masks provides the key mechanism to selectively choose from the KB r elevant weights to solve each task. By using MARK, we achieve state of the art r esults in several popular benchmarks, surpassing the best performing methods in terms of average accuracy by over 10% on the 20-Split-MiniImageNet dataset, whil e achieving almost zero forgetfulness using 55% of the number of parameters. Fur thermore, an ablation study provides evidence that, indeed, MARK is learning reu sable knowledge that is selectively used by each task.
****
A sampling-based circuit for optimal decision making
Camille Rullán Buxó, Cristina Savin
Many features of human and animal behavior can be understood in the framework of Bayesian inference and optimal decision making, but the biological substrate of such processes is not fully understood. Neural sampling provides a flexible cod e for probabilistic inference in high dimensions and explains key features of se nsory responses under experimental manipulations of uncertainty. However, since it encodes uncertainty implicitly, across time and neurons, it remains unclear h ow such representations can be used for decision making. Here we propose a spiki ng network model that maps neural samples of a task-specific marginal distributi on into an instantaneous representation of uncertainty via a procedure inspired by online kernel density estimation, so that its output can be readily used for decision making. Our model is consistent with experimental results at the level of single neurons and populations, and makes predictions for how neural respons es and decisions could be modulated by uncertainty and prior biases. More genera lly, our work brings together conflicting perspectives on probabilistic brain co mputation.
****
Compressed Video Contrastive Learning
Yuqi Huo, Mingyu Ding, Haoyu Lu, Nanyi Fei, Zhiwu Lu, Ji-Rong Wen, Ping Luo
This work concerns self-supervised video representation learning (SSVRL), one to pic that has received much attention recently. Since videos are storage-intensiv e and contain a rich source of visual content, models designed for SSVRL are exp ected to be storage- and computation-efficient, as well as effective. However, m ost existing methods only focus on one of the two objectives, failing to conside r both at the same time. In this work, for the first time, the seemingly contrad ictory goals are simultaneously achieved by exploiting compressed videos and cap turing mutual information between two input streams. Specifically, a novel Motio n Vector based Cross Guidance Contrastive learning approach (MVCGC) is proposed. For storage and computation efficiency, we choose to directly decode RGB frames and motion vectors (that resemble low-resolution optical flows) from compressed videos on-the-fly. To enhance the representation ability of the motion vectors, hence the effectiveness of our method, we design a cross guidance contrastive l earning algorithm based on multi-instance InfoNCE loss, where motion vectors can take supervision signals from RGB frames and vice versa. Comprehensive experime

nts on two downstream tasks show that our MVCGC yields new state-of-the-art whil
e being significantly more efficient than its competitors.
****

Uniform-PAC Bounds for Reinforcement Learning with Linear Function Approximation
Jiafan He, Dongruo Zhou, Quanquan Gu
We study reinforcement learning (RL) with linear function approximation. Existin
g algorithms for this problem only have high-probability regret and/or Probably
Approximately Correct (PAC) sample complexity guarantees, which cannot guarantee
 the convergence to the optimal policy. In this paper, in order to overcome the
limitation of existing algorithms, we propose a new algorithm called FLUTE, whic
h enjoys uniform-PAC convergence to the optimal policy with high probability. Th
e uniform-PAC guarantee is the strongest possible guarantee for reinforcement le
arning in the literature, which can directly imply both PAC and high probability
 regret bounds, making our algorithm superior to all existing algorithms with li
near function approximation. At the core of our algorithm is a novel minimax val
ue function estimator and a multi-level partition scheme to select the training
samples from historical observations. Both of these techniques are new and of in
dependent interest.
****

Attention Bottlenecks for Multimodal Fusion
Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, Chen Sun
Humans perceive the world by concurrently processing and fusing high-dimensional
 inputs from multiple modalities such as vision and audio. Machine perception m
odels, in stark contrast, are typically modality-specific and optimised for unim
odal benchmarks.A common approach for building multimodal models is to simply co
mbine multiple of these modality-specific architectures using late-stage fusion
of final representations or predictions ('late-fusion').Instead, we introduce a
novel transformer based architecture that uses 'attention bottlenecks' for modal
ity fusion at multiple layers. Compared to traditional pairwise self-attention,
 these bottlenecks force information between different modalities to pass throug
h a small number of '`bottleneck' latent units, requiring the model to collate a
nd condense the most relevant information in each modality and only share what i
s necessary. We find that such a strategy improves fusion performance, at the sa
me time reducing computational cost. We conduct thorough ablation studies, and a
chieve state-of-the-art results on multiple audio-visual classification benchmar
ks including Audioset, Epic-Kitchens and VGGSound. All code and models will be r
eleased.
****

Convergence of adaptive algorithms for constrained weakly convex optimization
Ahmet Alacaoglu, Yura Malitsky, 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.
****

On the Convergence of Step Decay Step-Size for Stochastic Optimization
Xiaoyu Wang, Sindri Magnússon, Mikael Johansson
Requests for name changes in the electronic proceedings will be accepted with 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.
****

BernNet: Learning Arbitrary Graph Spectral Filters via Bernstein Approximation
Mingguo He, Zhewei Wei, zengfeng Huang, Hongteng 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.
****

Co-evolution Transformer for Protein Contact Prediction

He Zhang, Fusong Ju, Jianwei Zhu, Liang He, Bin Shao, Nanning Zheng, Tie-Yan Liu
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Unsupervised Foreground Extraction via Deep Region Competition
Peiyu Yu, Sirui Xie, Xiaojian Ma, Yixin Zhu, Ying Nian Wu, Song-Chun Zhu
We present Deep Region Competition (DRC), an algorithm designed to extract foreground objects from images in a fully unsupervised manner. Foreground extraction can be viewed as a special case of generic image segmentation that focuses on identifying and disentangling objects from the background. In this work, we rethink the foreground extraction by reconciling energy-based prior with generative image modeling in the form of Mixture of Experts (MoE), where we further introduce the learned pixel re-assignment as the essential inductive bias to capture the regularities of background regions. With this modeling, the foreground-background partition can be naturally found through Expectation-Maximization (EM). We show that the proposed method effectively exploits the interaction between the mixture components during the partitioning process, which closely connects to region competition, a seminal approach for generic image segmentation. Experiments demonstrate that DRC exhibits more competitive performances on complex real-world data and challenging multi-object scenes compared with prior methods. Moreover, we show empirically that DRC can potentially generalize to novel foreground objects even from categories unseen during training.
****
Leveraging Spatial and Temporal Correlations in Sparsified Mean Estimation
Divyansh Jhunjhunwala, Ankur Mallick, Advait Gadhikar, Swanand Kadhe, Gauri Joshi
We study the problem of estimating at a central server the mean of a set of vectors distributed across several nodes (one vector per node). When the vectors are high-dimensional, the communication cost of sending entire vectors may be prohibitive, and it may be imperative for them to use sparsification techniques. While most existing work on sparsified mean estimation is agnostic to the characteristics of the data vectors, in many practical applications such as federated learning, there may be spatial correlations (similarities in the vectors sent by different nodes) or temporal correlations (similarities in the data sent by a single node over different iterations of the algorithm) in the data vectors. We leverage these correlations by simply modifying the decoding method used by the server to estimate the mean. We provide an analysis of the resulting estimation error as well as experiments for PCA, K-Means and Logistic Regression, which show that our estimators consistently outperform more sophisticated and expensive sparsification methods.
****
Last-iterate Convergence in Extensive-Form Games
Chung-Wei Lee, Christian Kroer, Haipeng Luo
Regret-based algorithms are highly efficient at finding approximate Nash equilibria in sequential games such as poker games. However, most regret-based algorithms, including counterfactual regret minimization (CFR) and its variants, rely on iterate averaging to achieve convergence. Inspired by recent advances on last-iterate convergence of optimistic algorithms in zero-sum normal-form games, we study this phenomenon in sequential games, and provide a comprehensive study of last-iterate convergence for zero-sum extensive-form games with perfect recall (EFGs), using various optimistic regret-minimization algorithms over treeplexes. This includes algorithms using the vanilla entropy or squared Euclidean norm regularizers, as well as their dilated versions which admit more efficient implementation. In contrast to CFR, we show that all of these algorithms enjoy last-iterate convergence, with some of them even converging exponentially fast. We also provide experiments to further support our theoretical results.
****
Class-Incremental Learning via Dual Augmentation

Fei Zhu, Zhen Cheng, Xu-yao Zhang, Cheng-lin Liu
Deep learning systems typically suffer from catastrophic forgetting of past knowledge when acquiring new skills continually. In this paper, we emphasize two dilemmas, representation bias and classifier bias in class-incremental learning, and present a simple and novel approach that employs explicit class augmentation (classAug) and implicit semantic augmentation (semanAug) to address the two biases, respectively. On the one hand, we propose to address the representation bias by learning transferable and diverse representations. Specifically, we investigate the feature representations in incremental learning based on spectral analysis and present a simple technique called classAug, to let the model see more classes during training for learning representations transferable across classes. On the other hand, to overcome the classifier bias, semanAug implicitly involves the simultaneous generating of an infinite number of instances of old classes in the deep feature space, which poses tighter constraints to maintain the decision boundary of previously learned classes. Without storing any old samples, our method can perform comparably with representative data replay based approaches.
****
Robust and Fully-Dynamic Coreset for Continuous-and-Bounded Learning (With Outliers) Problems
Zixiu Wang, Yiwen Guo, Hu Ding
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 and Reweighting the Univariate Losses for Multi-Label Ranking: Consistency and Generalization
Guoqiang Wu, Chongxuan LI, Kun Xu, 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-authors prior to requesting a name change in the electronic proceedings.
****
Fair Clustering Under a Bounded Cost
Seyed Esmaeili, Brian Brubach, Aravind Srinivasan, John Dickerson
Clustering is a fundamental unsupervised learning problem where a dataset is partitioned into clusters that consist of nearby points in a metric space. A recent variant, fair clustering, associates a color with each point representing its group membership and requires that each color has (approximately) equal representation in each cluster to satisfy group fairness. In this model, the cost of the clustering objective increases due to enforcing fairness in the algorithm. The relative increase in the cost, the ```````''price of fairness,'' can indeed be unbounded. Therefore, in this paper we propose to treat an upper bound on the clustering objective as a constraint on the clustering problem, and to maximize equality of representation subject to it. We consider two fairness objectives: the group utilitarian objective and the group egalitarian objective, as well as the group leximin objective which generalizes the group egalitarian objective. We derive fundamental lower bounds on the approximation of the utilitarian and egalitarian objectives and introduce algorithms with provable guarantees for them. For the leximin objective we introduce an effective heuristic algorithm. We further derive impossibility results for other natural fairness objectives. We conclude with experimental results on real-world datasets that demonstrate the validity of our algorithms.
****
Improving Calibration through the Relationship with Adversarial Robustness
Yao Qin, Xuezhi Wang, Alex Beutel, Ed Chi
Neural networks lack adversarial robustness, i.e., they are vulnerable to adversarial examples that through small perturbations to inputs cause incorrect predictions. Further, trust is undermined when models give miscalibrated predictions, i.e., the predicted probability is not a good indicator of how much we should t

rust our model. In this paper, we study the connection between adversarial robustness and calibration and find that the inputs for which the model is sensitive to small perturbations (are easily attacked) are more likely to have poorly calibrated predictions. Based on this insight, we examine if calibration can be improved by addressing those adversarially unrobust inputs. To this end, we propose Adversarial Robustness based Adaptive Label Smoothing (AR-AdaLS) that integrates the correlations of adversarial robustness and calibration into training by adaptively softening labels for an example based on how easily it can be attacked by an adversary. We find that our method, taking the adversarial robustness of the in-distribution data into consideration, leads to better calibration over the model even under distributional shifts. In addition, AR-AdaLS can also be applied to an ensemble model to further improve model calibration.
****
Credal Self-Supervised Learning
Julian Lienen, Eyke Hüllermeier
Self-training is an effective approach to semi-supervised learning. The key idea is to let the learner itself iteratively generate "pseudo-supervision" for unlabeled instances based on its current hypothesis. In combination with consistency regularization, pseudo-labeling has shown promising performance in various domains, for example in computer vision. To account for the hypothetical nature of the pseudo-labels, these are commonly provided in the form of probability distributions. Still, one may argue that even a probability distribution represents an excessive level of informedness, as it suggests that the learner precisely knows the ground-truth conditional probabilities. In our approach, we therefore allow the learner to label instances in the form of credal sets, that is, sets of (candidate) probability distributions. Thanks to this increased expressiveness, the learner is able to represent uncertainty and a lack of knowledge in a more flexible and more faithful manner. To learn from weakly labeled data of that kind, we leverage methods that have recently been proposed in the realm of so-called superset learning. In an exhaustive empirical evaluation, we compare our methodology to state-of-the-art self-supervision approaches, showing competitive to superior performance especially in low-label scenarios incorporating a high degree of uncertainty.
****
Spot the Difference: Detection of Topological Changes via Geometric Alignment
Per Steffen Czolbe, Aasa Feragen, Oswin Krause
Geometric alignment appears in a variety of applications, ranging from domain adaptation, optimal transport, and normalizing flows in machine learning; optical flow and learned augmentation in computer vision and deformable registration within biomedical imaging. A recurring challenge is the alignment of domains whose topology is not the same; a problem that is routinely ignored, potentially introducing bias in downstream analysis. As a first step towards solving such alignment problems, we propose an unsupervised algorithm for the detection of changes in image topology. The model is based on a conditional variational auto-encoder and detects topological changes between two images during the registration step. We account for both topological changes in the image under spatial variation and unexpected transformations. Our approach is validated on two tasks and datasets: detection of topological changes in microscopy images of cells, and unsupervised anomaly detection brain imaging.
****
Rethinking the Variational Interpretation of Accelerated Optimization Methods
Peiyuan Zhang, Antonio Orvieto, Hadi Daneshmand
The continuous-time model of Nesterov's momentum provides a thought-provoking perspective for understanding the nature of the acceleration phenomenon in convex optimization. One of the main ideas in this line of research comes from the field of classical mechanics and proposes to link Nesterov's trajectory to the solution of a set of Euler-Lagrange equations relative to the so-called Bregman Lagrangian. In the last years, this approach led to the discovery of many new (stochastic) accelerated algorithms and provided a solid theoretical foundation for the design of structure-preserving accelerated methods. In this work, we revisit th

is idea and provide an in-depth analysis of the action relative to the Bregman Lagrangian from the point of view of calculus of variations. Our main finding is that, while Nesterov's method is a stationary point for the action, it is often not a minimizer but instead a saddle point for this functional in the space of differentiable curves. This finding challenges the main intuition behind the variational interpretation of Nesterov's method and provides additional insights into the intriguing geometry of accelerated paths.
****
Linear and Kernel Classification in the Streaming Model: Improved Bounds for Heavy Hitters
Arvind Mahankali, 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.
****
A PAC-Bayes Analysis of Adversarial Robustness
Paul Viallard, Eric Guillaume VIDOT, Amaury Habrard, Emilie Morvant
We propose the first general PAC-Bayesian generalization bounds for adversarial robustness, that estimate, at test time, how much a model will be invariant to imperceptible perturbations in the input. Instead of deriving a worst-case analysis of the risk of a hypothesis over all the possible perturbations, we leverage the PAC-Bayesian framework to bound the averaged risk on the perturbations for majority votes (over the whole class of hypotheses). Our theoretically founded analysis has the advantage to provide general bounds (i) that are valid for any kind of attacks (i.e., the adversarial attacks), (ii) that are tight thanks to the PAC-Bayesian framework, (iii) that can be directly minimized during the learning phase to obtain a robust model on different attacks at test time.
****
SE(3)-equivariant prediction of molecular wavefunctions and electronic densities
Oliver Unke, Mihail Bogojeski, Michael Gastegger, Mario Geiger, Tess Smidt, Klaus-Robert Müller
Machine learning has enabled the prediction of quantum chemical properties with high accuracy and efficiency, allowing to bypass computationally costly ab initio calculations. Instead of training on a fixed set of properties, more recent approaches attempt to learn the electronic wavefunction (or density) as a central quantity of atomistic systems, from which all other observables can be derived. This is complicated by the fact that wavefunctions transform non-trivially under molecular rotations, which makes them a challenging prediction target. To solve this issue, we introduce general SE(3)-equivariant operations and building blocks for constructing deep learning architectures for geometric point cloud data and apply them to reconstruct wavefunctions of atomistic systems with unprecedented accuracy. Our model achieves speedups of over three orders of magnitude compared to ab initio methods and reduces prediction errors by up to two orders of magnitude compared to the previous state-of-the-art. This accuracy makes it possible to derive properties such as energies and forces directly from the wavefunction in an end-to-end manner. We demonstrate the potential of our approach in a transfer learning application, where a model trained on low accuracy reference wavefunctions implicitly learns to correct for electronic many-body interactions from observables computed at a higher level of theory. Such machine-learned wavefunction surrogates pave the way towards novel semi-empirical methods, offering resolution at an electronic level while drastically decreasing computational cost. Additionally, the predicted wavefunctions can serve as initial guess in conventional ab initio methods, decreasing the number of iterations required to arrive at a converged solution, thus leading to significant speedups without any loss of accuracy or robustness. While we focus on physics applications in this contribution, the proposed equivariant framework for deep learning on point clouds is promising also beyond, say, in computer vision or graphics.
****
Modified Frank Wolfe in Probability Space

Carson Kent, Jiajin Li, Jose Blanchet, Peter W Glynn
We propose a novel Frank-Wolfe (FW) procedure for the optimization of infinite-dimensional functionals of probability measures - a task which arises naturally in a wide range of areas including statistical learning (e.g. variational inference) and artificial intelligence (e.g. generative adversarial networks). Our FW procedure takes advantage of Wasserstein gradient flows and strong duality results recently developed in Distributionally Robust Optimization so that gradient steps (in the Wasserstein space) can be efficiently computed using finite-dimensional, convex optimization methods. We show how to choose the step sizes in order to guarantee exponentially fast iteration convergence, under mild assumptions on the functional to optimize. We apply our algorithm to a range of functionals arising from applications in nonparametric estimation.
****
Bayesian Optimization of Function Networks
Raul Astudillo, Peter Frazier
We consider Bayesian optimization of the output of a network of functions, where each function takes as input the output of its parent nodes, and where the network takes significant time to evaluate. Such problems arise, for example, in reinforcement learning, engineering design, and manufacturing. While the standard Bayesian optimization approach observes only the final output, our approach delivers greater query efficiency by leveraging information that the former ignores: intermediate output within the network. This is achieved by modeling the nodes of the network using Gaussian processes and choosing the points to evaluate using, as our acquisition function, the expected improvement computed with respect to the implied posterior on the objective. Although the non-Gaussian nature of this posterior prevents computing our acquisition function in closed form, we show that it can be efficiently maximized via sample average approximation. In addition, we prove that our method is asymptotically consistent, meaning that it finds a globally optimal solution as the number of evaluations grows to infinity, thus generalizing previously known convergence results for the expected improvement. Notably, this holds even though our method might not evaluate the domain densely, instead leveraging problem structure to leave regions unexplored. Finally, we show that our approach dramatically outperforms standard Bayesian optimization methods in several synthetic and real-world problems.
****
Look at What I’m Doing: Self-Supervised Spatial Grounding of Narrations in Instructional Videos
Reuben Tan, Bryan Plummer, Kate Saenko, Hailin Jin, Bryan Russell
We introduce the task of spatially localizing narrated interactions in videos. Key to our approach is the ability to learn to spatially localize interactions with self-supervision on a large corpus of videos with accompanying transcribed narrations. To achieve this goal, we propose a multilayer cross-modal attention network that enables effective optimization of a contrastive loss during training. We introduce a divided strategy that alternates between computing inter- and intra-modal attention across the visual and natural language modalities, which allows effective training via directly contrasting the two modalities' representations. We demonstrate the effectiveness of our approach by self-training on the HowTo100M instructional video dataset and evaluating on a newly collected dataset of localized described interactions in the YouCook2 dataset. We show that our approach outperforms alternative baselines, including shallow co-attention and full cross-modal attention. We also apply our approach to grounding phrases in images with weak supervision on Flickr30K and show that stacking multiple attention layers is effective and, when combined with a word-to-region loss, achieves state of the art on recall-at-one and pointing hand accuracies.
****
RETRIEVE: Coreset Selection for Efficient and Robust Semi-Supervised Learning
Krishnateja Killamsetty, Xujiang Zhao, Feng Chen, Rishabh Iyer
Requests for name changes in the electronic proceedings will be accepted with 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.
****
Collaborating with Humans without Human Data
DJ Strouse, Kevin McKee, Matt Botvinick, Edward Hughes, Richard Everett
Collaborating with humans requires rapidly adapting to their individual strengths, weaknesses, and preferences. Unfortunately, most standard multi-agent reinforcement learning techniques, such as self-play (SP) or population play (PP), produce agents that overfit to their training partners and do not generalize well to humans. Alternatively, researchers can collect human data, train a human model using behavioral cloning, and then use that model to train "human-aware" agents ("behavioral cloning play", or BCP). While such an approach can improve the generalization of agents to new human co-players, it involves the onerous and expensive step of collecting large amounts of human data first. Here, we study the problem of how to train agents that collaborate well with human partners without using human data. We argue that the crux of the problem is to produce a diverse set of training partners. Drawing inspiration from successful multi-agent approaches in competitive domains, we find that a surprisingly simple approach is highly effective. We train our agent partner as the best response to a population of self-play agents and their past checkpoints taken throughout training, a method we call Fictitious Co-Play (FCP). Our experiments focus on a two-player collaborative cooking simulator that has recently been proposed as a challenge problem for coordination with humans. We find that FCP agents score significantly higher than SP, PP, and BCP when paired with novel agent and human partners. Furthermore, humans also report a strong subjective preference to partnering with FCP agents over all baselines.
****
Training Feedback Spiking Neural Networks by Implicit Differentiation on the Equilibrium State
Mingqing Xiao, Qingyan Meng, Zongpeng Zhang, Yisen Wang, Zhouchen Lin
Spiking neural networks (SNNs) are brain-inspired models that enable energy-efficient implementation on neuromorphic hardware. However, the supervised training of SNNs remains a hard problem due to the discontinuity of the spiking neuron model. Most existing methods imitate the backpropagation framework and feedforward architectures for artificial neural networks, and use surrogate derivatives or compute gradients with respect to the spiking time to deal with the problem. These approaches either accumulate approximation errors or only propagate information limitedly through existing spikes, and usually require information propagation along time steps with large memory costs and biological implausibility. In this work, we consider feedback spiking neural networks, which are more brain-like, and propose a novel training method that does not rely on the exact reverse of the forward computation. First, we show that the average firing rates of SNNs with feedback connections would gradually evolve to an equilibrium state along time, which follows a fixed-point equation. Then by viewing the forward computation of feedback SNNs as a black-box solver for this equation, and leveraging the implicit differentiation on the equation, we can compute the gradient for parameters without considering the exact forward procedure. In this way, the forward and backward procedures are decoupled and therefore the problem of non-differentiable spiking functions is avoided. We also briefly discuss the biological plausibility of implicit differentiation, which only requires computing another equilibrium. Extensive experiments on MNIST, Fashion-MNIST, N-MNIST, CIFAR-10, and CIFAR-100 demonstrate the superior performance of our method for feedback models with fewer neurons and parameters in a small number of time steps. Our code is available at https://github.com/pkuxmq/IDE-FSNN.
****
Online Selective Classification with Limited Feedback
Aditya Gangrade, Anil Kag, Ashok Cutkosky, Venkatesh Saligrama
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****
Controlled Text Generation as Continuous Optimization with Multiple Constraints
Sachin Kumar, Eric Malmi, Aliaksei Severyn, Yulia Tsvetkov
As large-scale language model pretraining pushes the state-of-the-art in text ge
neration, recent work has turned to controlling attributes of the text such mode
ls generate. While modifying the pretrained models via fine-tuning remains the p
opular approach, it incurs a significant computational cost and can be infeasibl
e due to a lack of appropriate data. As an alternative, we propose \textsc{MuCoC
O}---a flexible and modular algorithm for controllable inference from pretrained
 models. We formulate the decoding process as an optimization problem that allow
s for multiple attributes we aim to control to be easily incorporated as differe
ntiable constraints. By relaxing this discrete optimization to a continuous one,
 we make use of Lagrangian multipliers and gradient-descent-based techniques to
generate the desired text. We evaluate our approach on controllable machine tran
slation and style transfer with multiple sentence-level attributes and observe s
ignificant improvements over baselines.
****
S^3: Sign-Sparse-Shift Reparametrization for Effective Training of Low-bit Shi
ft Networks
Xinlin Li, Bang Liu, Yaoliang Yu, Wulong Liu, Chunjing XU, Vahid Partovi Nia
Requests for name changes in the electronic proceedings will be accepted with 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.
****
Implicit MLE: Backpropagating Through Discrete Exponential Family Distributions
Mathias Niepert, Pasquale Minervini, Luca Franceschi
Combining discrete probability distributions and combinatorial optimization prob
lems with neural network components has numerous applications but poses several
challenges. We propose Implicit Maximum Likelihood Estimation (I-MLE), a framewo
rk for end-to-end learning of models combining discrete exponential family distr
ibutions and differentiable neural components. I-MLE is widely applicable as it
only requires the ability to compute the most probable states and does not rely
on smooth relaxations. The framework encompasses several approaches such as pert
urbation-based implicit differentiation and recent methods to differentiate thro
ugh black-box combinatorial solvers. We introduce a novel class of noise distrib
utions for approximating marginals via perturb-and-MAP. Moreover, we show that I
-MLE simplifies to maximum likelihood estimation when used in some recently stud
ied learning settings that involve combinatorial solvers. Experiments on several
 datasets suggest that I-MLE is competitive with and often outperforms existing
approaches which rely on problem-specific relaxations.
****
Scaling up Continuous-Time Markov Chains Helps Resolve Underspecification
Alkis Gotovos, Rebekka Burkholz, John Quackenbush, Stefanie Jegelka
Modeling the time evolution of discrete sets of items (e.g., genetic mutations)
is a fundamental problem in many biomedical applications. We approach this probl
em through the lens of continuous-time Markov chains, and show that the resultin
g learning task is generally underspecified in the usual setting of cross-sectio
nal data. We explore a perhaps surprising remedy: including a number of addition
al independent items can help determine time order, and hence resolve underspeci
fication. This is in sharp contrast to the common practice of limiting the analy
sis to a small subset of relevant items, which is followed largely due to poor s
caling of existing methods. To put our theoretical insight into practice, we dev
elop an approximate likelihood maximization method for learning continuous-time
Markov chains, which can scale to hundreds of items and is orders of magnitude f
aster than previous methods. We demonstrate the effectiveness of our approach on
 synthetic and real cancer data.
****
Do Neural Optimal Transport Solvers Work? A Continuous Wasserstein-2 Benchmark
Alexander Korotin, Lingxiao Li, Aude Genevay, Justin M. Solomon, Alexander Filip

pov, Evgeny Burnaev
Despite the recent popularity of neural network-based solvers for optimal transport (OT), there is no standard quantitative way to evaluate their performance. In this paper, we address this issue for quadratic-cost transport---specifically, computation of the Wasserstein-2 distance, a commonly-used formulation of optimal transport in machine learning. To overcome the challenge of computing ground truth transport maps between continuous measures needed to assess these solvers, we use input-convex neural networks (ICNN) to construct pairs of measures whose ground truth OT maps can be obtained analytically. This strategy yields pairs of continuous benchmark measures in high-dimensional spaces such as spaces of images. We thoroughly evaluate existing optimal transport solvers using these benchmark measures. Even though these solvers perform well in downstream tasks, many do not faithfully recover optimal transport maps. To investigate the cause of this discrepancy, we further test the solvers in a setting of image generation. Our study reveals crucial limitations of existing solvers and shows that increased OT accuracy does not necessarily correlate to better results downstream.
****
Linear Convergence in Federated Learning: Tackling Client Heterogeneity and Sparse Gradients
Aritra Mitra, Rayana Jaafar, George J. Pappas, Hamed Hassani
We consider a standard federated learning (FL) setup where a group of clients periodically coordinate with a central server to train a statistical model. We develop a general algorithmic framework called FedLin to tackle some of the key challenges intrinsic to FL, namely objective heterogeneity, systems heterogeneity, and infrequent and imprecise communication. Our framework is motivated by the observation that under these challenges, various existing FL algorithms suffer from a fundamental speed-accuracy conflict: they either guarantee linear convergence but to an incorrect point, or convergence to the global minimum but at a sub-linear rate, i.e., fast convergence comes at the expense of accuracy. In contrast, when the clients' local loss functions are smooth and strongly convex, we show that FedLin guarantees linear convergence to the global minimum, despite arbitrary objective and systems heterogeneity. We then establish matching upper and lower bounds on the convergence rate of FedLin that highlight the effects of infrequent, periodic communication. Finally, we show that FedLin preserves linear convergence rates under aggressive gradient sparsification, and quantify the effect of the compression level on the convergence rate. Notably, our work is the first to provide tight linear convergence rate guarantees, and constitutes the first comprehensive analysis of gradient sparsification in FL.
****
On the Convergence of Prior-Guided Zeroth-Order Optimization Algorithms
Shuyu Cheng, Guoqiang Wu, Jun Zhu
Zeroth-order (ZO) optimization is widely used to handle challenging tasks, such as query-based black-box adversarial attacks and reinforcement learning. Various attempts have been made to integrate prior information into the gradient estimation procedure based on finite differences, with promising empirical results. However, their convergence properties are not well understood. This paper makes an attempt to fill up this gap by analyzing the convergence of prior-guided ZO algorithms under a greedy descent framework with various gradient estimators. We provide a convergence guarantee for the prior-guided random gradient-free (PRGF) algorithms. Moreover, to further accelerate over greedy descent methods, we present a new accelerated random search (ARS) algorithm that incorporates prior information, together with a convergence analysis. Finally, our theoretical results are confirmed by experiments on several numerical benchmarks as well as adversarial attacks.
****
Revisit Multimodal Meta-Learning through the Lens of Multi-Task Learning
Milad Abdollahzadeh, Touba Malekzadeh, Ngai-Man (Man) Cheung
Multimodal meta-learning is a recent problem that extends conventional few-shot meta-learning by generalizing its setup to diverse multimodal task distributions. This setup makes a step towards mimicking how humans make use of a diverse set

of prior skills to learn new skills. Previous work has achieved encouraging per formance. In particular, in spite of the diversity of the multimodal tasks, prev ious work claims that a single meta-learner trained on a multimodal distribution can sometimes outperform multiple specialized meta-learners trained on individu al unimodal distributions. The improvement is attributed to knowledge transfer b etween different modes of task distributions. However, there is no deep investig ation to verify and understand the knowledge transfer between multimodal tasks. Our work makes two contributions to multimodal meta-learning. First, we propose a method to quantify knowledge transfer between tasks of different modes at a mi cro-level. Our quantitative, task-level analysis is inspired by the recent trans ference idea from multi-task learning. Second, inspired by hard parameter sharin g in multi-task learning and a new interpretation of related work, we propose a new multimodal meta-learner that outperforms existing work by considerable margi ns. While the major focus is on multimodal meta-learning, our work also attempts to shed light on task interaction in conventional meta-learning. The code for t his project is available at https://miladabd.github.io/KML.
****
Dynamic Sasvi: Strong Safe Screening for Norm-Regularized Least Squares
Hiroaki Yamada, Makoto Yamada
A recently introduced technique, called safe screening,'' for a sparse optimizat ion problem allows us to identify irrelevant variables in the early stages of op timization. In this paper, we first propose a flexible framework for safe screen ing based on the Fenchel--Rockafellar duality and then derive a strong safe scre ening rule for norm-regularized least squares using the proposed framework. We r efer to the proposed screening rule for norm-regularized least squares asdynamic Sasvi'' because it can be interpreted as a generalization of Sasvi. Unlike the original Sasvi, it does not require the exact solution of a more strongly regula rized problem; hence, it works safely in practice. We show that our screening ru le always eliminates more features compared with the existing state-of-the-art m ethods.
****
What Matters for Adversarial Imitation Learning?
Manu Orsini, Anton Raichuk, Leonard Hussenot, Damien Vincent, Robert Dadashi, Se rtan Girgin, Matthieu Geist, Olivier Bachem, Olivier Pietquin, Marcin Andrychowi cz
Adversarial imitation learning has become a popular framework for imitation in c ontinuous control. Over the years, several variations of its components were pro posed to enhance the performance of the learned policies as well as the sample c omplexity of the algorithm. In practice, these choices are rarely tested all tog ether in rigorous empirical studies.It is therefore difficult to discuss and und erstand what choices, among the high-level algorithmic options as well as low-l evel implementation details, matter. To tackle this issue, we implement more tha n 50 of these choices in a generic adversarial imitation learning frameworkand i nvestigate their impacts in a large-scale study (>500k trained agents) with both synthetic and human-generated demonstrations. We analyze the key results and hi ghlight the most surprising findings.
****
Sequential Causal Imitation Learning with Unobserved Confounders
Daniel Kumor, Junzhe Zhang, Elias Bareinboim
"Monkey see monkey do" is an age-old adage, referring to naive imitation without a deep understanding of a system's underlying mechanics. Indeed, if a demonstra tor has access to information unavailable to the imitator (monkey), such as a di fferent set of sensors, then no matter how perfectly the imitator models its per ceived environment (See), attempting to directly reproduce the demonstrator's be havior (Do) can lead to poor outcomes. Imitation learning in the presence of a m ismatch between demonstrator and imitator has been studied in the literature und er the rubric of causal imitation learning (Zhang et. al. 2020), but existing s olutions are limited to single-stage decision-making. This paper investigates th e problem of causal imitation learning in sequential settings, where the imitato r must make multiple decisions per episode. We develop a graphical criterion tha

t is both necessary and sufficient for determining the feasibility of causal imitation, providing conditions when an imitator can match a demonstrator's performance despite differing capabilities. Finally, we provide an efficient algorithm for determining imitability, and corroborate our theory with simulations.
****

Topic Modeling Revisited: A Document Graph-based Neural Network Perspective
Dazhong Shen, Chuan Qin, Chao Wang, Zheng Dong, Hengshu Zhu, Hui Xiong
Most topic modeling approaches are based on the bag-of-words assumption, where each word is required to be conditionally independent in the same document. As a result, both of the generative story and the topic formulation have totally ignored the semantic dependency among words, which is important for improving the semantic comprehension and model interpretability. To this end, in this paper, we revisit the task of topic modeling by transforming each document into a directed graph with word dependency as edges between word nodes, and develop a novel approach, namely Graph Neural Topic Model (GNTM). Specifically, in GNTM, a well-defined probabilistic generative story is designed to model both the graph structure and word sets with multinomial distributions on the vocabulary and word dependency edge set as the topics. Meanwhile, a Neural Variational Inference (NVI) approach is proposed to learn our model with graph neural networks to encode the document graphs. Besides, we theoretically demonstrate that Latent Dirichlet Allocation (LDA) can be derived from GNTM as a special case with similar objective functions. Finally, extensive experiments on four benchmark datasets have clearly demonstrated the effectiveness and interpretability of GNTM compared with state-of-the-art baselines.
****

Hard-Attention for Scalable Image Classification
Athanasios Papadopoulos, Pawel Korus, Nasir Memon
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 Routing under Uncertainty: Adaptive Learning in Congestion Games via Exponential Weights
Dong Quan Vu, Kimon Antonakopoulos, Panayotis Mertikopoulos
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

Profiling Pareto Front With Multi-Objective Stein Variational Gradient Descent
Xingchao Liu, Xin Tong, Qiang Liu
Finding diverse and representative Pareto solutions from the Pareto front is a key challenge in multi-objective optimization (MOO). In this work, we propose a novel gradient-based algorithm for profiling Pareto front by using Stein variational gradient descent (SVGD). We also provide a counterpart of our method based on Langevin dynamics. Our methods iteratively update a set of points in a parallel fashion to push them towards the Pareto front using multiple gradient descent, while encouraging the diversity between the particles by using the repulsive force mechanism in SVGD, or diffusion noise in Langevin dynamics. Compared with existing gradient-based methods that require predefined preference functions, our method can work efficiently in high dimensional problems, and can obtain more diverse solutions evenly distributed in the Pareto front. Moreover, our methods are theoretically guaranteed to converge to the Pareto front. We demonstrate the effectiveness of our method, especially the SVGD algorithm, through extensive experiments, showing its superiority over existing gradient-based algorithms.
****

MAP Propagation Algorithm: Faster Learning with a Team of Reinforcement Learning Agents
Stephen Chung

Nearly all state-of-the-art deep learning algorithms rely on error backpropagati
on, which is generally regarded as biologically implausible. An alternative way
of training an artificial neural network is through treating each unit in the ne
twork as a reinforcement learning agent, and thus the network is considered as a
 team of agents. As such, all units can be trained by REINFORCE, a local learnin
g rule modulated by a global signal that is more consistent with biologically ob
served forms of synaptic plasticity. Although this learning rule follows the gra
dient of return in expectation, it suffers from high variance and thus the low s
peed of learning, rendering it impractical to train deep networks. We therefore
propose a novel algorithm called MAP propagation to reduce this variance signifi
cantly while retaining the local property of the learning rule. Experiments demo
nstrated that MAP propagation could solve common reinforcement learning tasks at
 a similar speed to backpropagation when applied to an actor-critic network. Our
 work thus allows for the broader application of teams of agents in deep reinfor
cement learning.
****
TransGAN: Two Pure Transformers Can Make One Strong GAN, and That Can Scale Up
Yifan Jiang, Shiyu Chang, Zhangyang 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.
****
A Central Limit Theorem for Differentially Private Query Answering
Jinshuo Dong, Weijie Su, Linjun 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.
****
Differential Privacy Dynamics of Langevin Diffusion and Noisy Gradient Descent
Rishav Chourasia, Jiayuan Ye, Reza Shokri
What is the information leakage of an iterative randomized learning algorithm ab
out its training data, when the internal state of the algorithm is \emph{private
}? How much is the contribution of each specific training epoch to the informati
on leakage through the released model? We study this problem for noisy gradient
descent algorithms, and model the \emph{dynamics} of R\'enyi differential privac
y loss throughout the training process. Our analysis traces a provably \emph{ti
ght} bound on the R\'enyi divergence between the pair of probability distributio
ns over parameters of models trained on neighboring datasets. We prove that the
 privacy loss converges exponentially fast, for smooth and strongly convex loss
functions, which is a significant improvement over composition theorems (which o
ver-estimate the privacy loss by upper-bounding its total value over all interme
diate gradient computations). For Lipschitz, smooth, and strongly convex loss fu
nctions, we prove optimal utility with a small gradient complexity for noisy gra
dient descent algorithms.
****
Data driven semi-supervised learning
Maria-Florina F. Balcan, Dravyansh Sharma
We consider a novel data driven approach for designing semi-supervised learning
algorithms that can effectively learn with only a small number of labeled exampl
es. We focus on graph-based techniques, where the unlabeled examples are connect
ed in a graph under the implicit assumption that similar nodes likely have simi
lar labels. Over the past two decades, several elegant graph-based semi-supervis
ed learning algorithms for inferring the labels of the unlabeled examples given
the graph and a few labeled examples have been proposed. However, the problem of
 how to create the graph (which impacts the practical usefulness of these method
s significantly) has been relegated to heuristics and domain-specific art, and n
o general principles have been proposed. In this work we present a novel data d
riven approach for learning the graph and provide strong formal guarantees in bo

th the distributional and online learning formalizations. We show how to leverag
e problem instances coming from an underlying problem domain to learn the graph
hyperparameters for commonly used parametric families of graphs that provably pe
rform well on new instances from the same domain. We obtain low regret and effic
ient algorithms in the online setting, and generalization guarantees in the dist
ributional setting. We also show how to combine several very different similarit
y metrics and learn multiple hyperparameters, our results hold for large classe
s of problems. We expect some of the tools and techniques we develop along the w
ay to be of independent interest, for data driven algorithms more generally.
****

Online Meta-Learning via Learning with Layer-Distributed Memory
Sudarshan Babu, Pedro Savarese, Michael Maire
We demonstrate that efficient meta-learning can be achieved via end-to-end train
ing of deep neural networks with memory distributed across layers. The persiste
nt state of this memory assumes the entire burden of guiding task adaptation. M
oreover, its distributed nature is instrumental in orchestrating adaptation. Ab
lation experiments demonstrate that providing relevant feedback to memory units
distributed across the depth of the network enables them to guide adaptation thr
oughout the entire network. Our results show that this is a successful strategy
 for simplifying meta-learning -- often cast as a bi-level optimization problem
-- to standard end-to-end training, while outperforming gradient-based, prototyp
e-based, and other memory-based meta-learning strategies. Additionally, our ada
ptation strategy naturally handles online learning scenarios with a significant
delay between observing a sample and its corresponding label -- a setting in whi
ch other approaches struggle. Adaptation via distributed memory is effective ac
ross a wide range of learning tasks, ranging from classification to online few-s
hot semantic segmentation.
****

Physics-Integrated Variational Autoencoders for Robust and Interpretable Generat
ive Modeling
Naoya Takeishi, Alexandros Kalousis
Integrating physics models within machine learning models holds considerable pro
mise toward learning robust models with improved interpretability and abilities
to extrapolate. In this work, we focus on the integration of incomplete physics
models into deep generative models. In particular, we introduce an architecture
of variational autoencoders (VAEs) in which a part of the latent space is ground
ed by physics. A key technical challenge is to strike a balance between the inco
mplete physics and trainable components such as neural networks for ensuring tha
t the physics part is used in a meaningful manner. To this end, we propose a reg
ularized learning method that controls the effect of the trainable components an
d preserves the semantics of the physics-based latent variables as intended. We
not only demonstrate generative performance improvements over a set of synthetic
 and real-world datasets, but we also show that we learn robust models that can
consistently extrapolate beyond the training distribution in a meaningful manner
. Moreover, we show that we can control the generative process in an interpretab
le manner.
****

Characterizing the risk of fairwashing
Ulrich Aïvodji, Hiromi Arai, Sébastien Gambs, Satoshi Hara
Fairwashing refers to the risk that an unfair black-box model can be explained b
y a fairer model through post-hoc explanation manipulation. In this paper, we in
vestigate the capability of fairwashing attacks by analyzing their fidelity-unfa
irness trade-offs. In particular, we show that fairwashed explanation models can
 generalize beyond the suing group (i.e., data points that are being explained),
 meaning that a fairwashed explainer can be used to rationalize subsequent unfai
r decisions of a black-box model. We also demonstrate that fairwashing attacks c
an transfer across black-box models, meaning that other black-box models can per
form fairwashing without explicitly using their predictions. This generalization
 and transferability of fairwashing attacks imply that their detection will be d
ifficult in practice. Finally, we propose an approach to quantify the risk of fa

irwashing, which is based on the computation of the range of the unfairness of high-fidelity explainers.
****
Qimera: Data-free Quantization with Synthetic Boundary Supporting Samples
Kanghyun Choi, Deokki Hong, Noseong Park, Youngsok Kim, Jinho Lee
Model quantization is known as a promising method to compress deep neural networks, especially for inferences on lightweight mobile or edge devices. However, model quantization usually requires access to the original training data to maintain the accuracy of the full-precision models, which is often infeasible in real-world scenarios for security and privacy issues.A popular approach to perform quantization without access to the original data is to use synthetically generated samples, based on batch-normalization statistics or adversarial learning.However, the drawback of such approaches is that they primarily rely on random noise input to the generator to attain diversity of the synthetic samples. We find that this is often insufficient to capture the distribution of the original data, especially around the decision boundaries.To this end, we propose Qimera, a method that uses superposed latent embeddings to generate synthetic boundary supporting samples.For the superposed embeddings to better reflect the original distribution, we also propose using an additional disentanglement mapping layer and extracting information from the full-precision model.The experimental results show that Qimera achieves state-of-the-art performances for various settings on data-free quantization. Code is available at https://github.com/iamkanghyunchoi/qimera.
****
Embedding Principle of Loss Landscape of Deep Neural Networks
Yaoyu Zhang, Zhongwang Zhang, Tao Luo, Zhiqin J Xu
Understanding the structure of loss landscape of deep neural networks (DNNs) is obviously important. In this work, we prove an embedding principle that the loss landscape of a DNN "contains" all the critical points of all the narrower DNNs. More precisely, we propose a critical embedding such that any critical point, e.g., local or global minima, of a narrower DNN can be embedded to a critical point/affine subspace of the target DNN with higher degeneracy and preserving the DNN output function. Note that, given any training data, differentiable loss function and differentiable activation function, this embedding structure of critical points holds.This general structure of DNNs is starkly different from other nonconvex problems such as protein-folding.Empirically, we find that a wide DNN is often attracted by highly-degenerate critical points that are embedded from narrow DNNs. The embedding principle provides a new perspective to study the general easy optimization of wide DNNs and unravels a potential implicit low-complexity regularization during the training.Overall, our work provides a skeleton for the study of loss landscape of DNNs and its implication, by which a more exact and comprehensive understanding can be anticipated in the near future.
****
Adversarial Reweighting for Partial Domain Adaptation
Xiang Gu, Xi Yu, yan yang, Jian Sun, Zongben Xu
Partial domain adaptation (PDA) has gained much attention due to its practical setting. The current PDA methods usually adapt the feature extractor by aligning the target and reweighted source domain distributions. In this paper, we experimentally find that the feature adaptation by the reweighted distribution alignment in some state-of-the-art PDA methods is not robust to the ``noisy'' weights of source domain data, leading to negative domain transfer on some challenging benchmarks. To tackle the challenge of negative domain transfer, we propose a novel Adversarial Reweighting (AR) approach that adversarially learns the weights of source domain data to align the source and target domain distributions, and the transferable deep recognition network is learned on the reweighted source domain data. Based on this idea, we propose a training algorithm that alternately updates the parameters of the network and optimizes the weights of source domain data. Extensive experiments show that our method achieves state-of-the-art results on the benchmarks of ImageNet-Caltech, Office-Home, VisDA-2017, and DomainNet. Ablation studies also confirm the effectiveness of our approach.
****

M-FAC: Efficient Matrix-Free Approximations of Second-Order Information
Elias Frantar, Eldar Kurtic, Dan Alistarh
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Graph Adversarial Self-Supervised Learning
Longqi Yang, Liangliang Zhang, Wenjing Yang
This paper studies a long-standing problem of learning the representations of a whole graph without human supervision. The recent self-supervised learning methods train models to be invariant to the transformations (views) of the inputs. However, designing these views requires the experience of human experts. Inspired by adversarial training, we propose an adversarial self-supervised learning (\texttt{GASSL}) framework for learning unsupervised representations of graph data without any handcrafted views. \texttt{GASSL} automatically generates challenging views by adding perturbations to the input and are adversarially trained with respect to the encoder. Our method optimizes the min-max problem and utilizes a gradient accumulation strategy to accelerate the training process. Experimental on ten graph classification datasets show that the proposed approach is superior to state-of-the-art self-supervised learning baselines, which are competitive with supervised models.
****
Anti-Backdoor Learning: Training Clean Models on Poisoned Data
Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, Xingjun Ma
Backdoor attack has emerged as a major security threat to deep neural networks (DNNs). While existing defense methods have demonstrated promising results on detecting or erasing backdoors, it is still not clear whether robust training methods can be devised to prevent the backdoor triggers being injected into the trained model in the first place. In this paper, we introduce the concept of \emph{anti-backdoor learning}, aiming to train \emph{clean} models given backdoor-poisoned data. We frame the overall learning process as a dual-task of learning the \emph{clean} and the \emph{backdoor} portions of data. From this view, we identify two inherent characteristics of backdoor attacks as their weaknesses: 1) the models learn backdoored data much faster than learning with clean data, and the stronger the attack the faster the model converges on backdoored data; 2) the backdoor task is tied to a specific class (the backdoor target class). Based on these two weaknesses, we propose a general learning scheme, Anti-Backdoor Learning (ABL), to automatically prevent backdoor attacks during training. ABL introduces a two-stage \emph{gradient ascent} mechanism for standard training to 1) help isolate backdoor examples at an early training stage, and 2) break the correlation between backdoor examples and the target class at a later training stage. Through extensive experiments on multiple benchmark datasets against 10 state-of-the-art attacks, we empirically show that ABL-trained models on backdoor-poisoned data achieve the same performance as they were trained on purely clean data. Code is available at \url{https://github.com/bboylyg/ABL}.
****
Locally Most Powerful Bayesian Test for Out-of-Distribution Detection using Deep Generative Models
Keunseo Kim, JunCheol Shin, Heeyoung Kim
Several out-of-distribution (OOD) detection scores have been recently proposed for deep generative models because the direct use of the likelihood threshold for OOD detection has been shown to be problematic. In this paper, we propose a new OOD score based on a Bayesian hypothesis test called the locally most powerful Bayesian test (LMPBT). The LMPBT is locally most powerful in that the alternative hypothesis (the representative parameter for the OOD sample) is specified to maximize the probability that the Bayes factor exceeds the evidence threshold in favor of the alternative hypothesis provided that the parameter specified under the alternative hypothesis is in the neighborhood of the parameter specified under the null hypothesis. That is, under this neighborhood parameter condition, th

e test with the proposed alternative hypothesis maximizes the probability of cor rect detection of OOD samples. We also propose numerical strategies for more eff icient and reliable computation of the LMPBT for practical application to deep g enerative models. Evaluations conducted of the OOD detection performance of the LMPBT on various benchmark datasets demonstrate its superior performance over ex isting OOD detection methods.
****
Stable Neural ODE with Lyapunov-Stable Equilibrium Points for Defending Against Adversarial Attacks
Qiyu Kang, Yang Song, Qinxu Ding, Wee Peng Tay
Deep neural networks (DNNs) are well-known to be vulnerable to adversarial attac ks, where malicious human-imperceptible perturbations are included in the input to the deep network to fool it into making a wrong classification. Recent studie s have demonstrated that neural Ordinary Differential Equations (ODEs) are intri nsically more robust against adversarial attacks compared to vanilla DNNs. In th is work, we propose a neural ODE with Lyapunov-stable equilibrium points for def ending against adversarial attacks (SODEF). By ensuring that the equilibrium poi nts of the ODE solution used as part of SODEF are Lyapunov-stable, the ODE solut ion for an input with a small perturbation converges to the same solution as the unperturbed input. We provide theoretical results that give insights into the s tability of SODEF as well as the choice of regularizers to ensure its stability. Our analysis suggests that our proposed regularizers force the extracted featur e points to be within a neighborhood of the Lyapunov-stable equilibrium points o f the SODEF ODE. SODEF is compatible with many defense methods and can be applie d to any neural network's final regressor layer to enhance its stability against adversarial attacks.
****
Robust Compressed Sensing MRI with Deep Generative Priors
Ajil Jalal, Marius Arvinte, Giannis Daras, Eric Price, Alexandros G. Dimakis, Jo n Tamir
The CSGM framework (Bora-Jalal-Price-Dimakis'17) has shown that deepgenerative p riors can be powerful tools for solving inverse problems.However, to date this f ramework has been empirically successful only oncertain datasets (for example, h uman faces and MNIST digits), and itis known to perform poorly on out-of-distrib ution samples. In thispaper, we present the first successful application of the CSGMframework on clinical MRI data. We train a generative prior on brainscans fr om the fastMRI dataset, and show that posterior sampling viaLangevin dynamics ac hieves high quality reconstructions. Furthermore,our experiments and theory show that posterior sampling is robust tochanges in the ground-truth distribution an d measurement process.Our code and models are available at: \url{https://github. com/utcsilab/csgm-mri-langevin}.
****
H-NeRF: Neural Radiance Fields for Rendering and Temporal Reconstruction of Huma ns in Motion
Hongyi Xu, Thiemo Alldieck, Cristian Sminchisescu
We present neural radiance fields for rendering and temporal (4D) reconstruction of humans in motion (H-NeRF), as captured by a sparse set of cameras or even fr om a monocular video. Our approach combines ideas from neural scene representati on, novel-view synthesis, and implicit statistical geometric human representatio ns, coupled using novel loss functions. Instead of learning a radiance field wit h a uniform occupancy prior, we constrain it by a structured implicit human body model, represented using signed distance functions. This allows us to robustly fuse information from sparse views and generalize well beyond the poses or views observed in training. Moreover, we apply geometric constraints to co-learn the structure of the observed subject -- including both body and clothing -- and to regularize the radiance field to geometrically plausible solutions. Extensive ex periments on multiple datasets demonstrate the robustness and the accuracy of ou r approach, its generalization capabilities significantly outside a small traini ng set of poses and views, and statistical extrapolation beyond the observed sha pe.

****
DOBF: A Deobfuscation Pre-Training Objective for Programming Languages
Marie-Anne Lachaux, Baptiste Roziere, Marc Szafraniec, Guillaume Lample
Recent advances in self-supervised learning have dramatically improved the state of the art on a wide variety of tasks. However, research in language model pre-training has mostly focused on natural languages, and it is unclear whether models like BERT and its variants provide the best pre-training when applied to other modalities, such as source code. In this paper, we introduce a new pre-training objective, DOBF, that leverages the structural aspect of programming languages and pre-trains a model to recover the original version of obfuscated source code. We show that models pre-trained with DOBF significantly outperform existing approaches on multiple downstream tasks, providing relative improvements of up to 12.2% in unsupervised code translation, and 5.3% in natural language code search. Incidentally, we found that our pre-trained model is able to deobfuscate fully obfuscated source files, and to suggest descriptive variable names.
****
Detecting Errors and Estimating Accuracy on Unlabeled Data with Self-training Ensembles
Jiefeng Chen, Frederick Liu, Besim Avci, Xi Wu, Yingyu Liang, Somesh Jha
When a deep learning model is deployed in the wild, it can encounter test data drawn from distributions different from the training data distribution and suffer drop in performance. For safe deployment, it is essential to estimate the accuracy of the pre-trained model on the test data. However, the labels for the test inputs are usually not immediately available in practice, and obtaining them can be expensive. This observation leads to two challenging tasks: (1) unsupervised accuracy estimation, which aims to estimate the accuracy of a pre-trained classifier on a set of unlabeled test inputs; (2) error detection, which aims to identify mis-classified test inputs. In this paper, we propose a principled and practically effective framework that simultaneously addresses the two tasks. The proposed framework iteratively learns an ensemble of models to identify mis-classified data points and performs self-training to improve the ensemble with the identified points. Theoretical analysis demonstrates that our framework enjoys provable guarantees for both accuracy estimation and error detection under mild conditions readily satisfied by practical deep learning models. Along with the framework, we proposed and experimented with two instantiations and achieved state-of-the-art results on 59 tasks. For example, on iWildCam, one instantiation reduces the estimation error for unsupervised accuracy estimation by at least 70% and improves the F1 score for error detection by at least 4.7% compared to existing methods.
****
Exploiting Chain Rule and Bayes' Theorem to Compare Probability Distributions
Huangjie Zheng, Mingyuan Zhou
To measure the difference between two probability distributions, referred to as the source and target, respectively, we exploit both the chain rule and Bayes' theorem to construct conditional transport (CT), which is constituted by both a forward component and a backward one. The forward CT is the expected cost of moving a source data point to a target one, with their joint distribution defined by the product of the source probability density function (PDF) and a source-dependent conditional distribution, which is related to the target PDF via Bayes' theorem. The backward CT is defined by reversing the direction. The CT cost can be approximated by replacing the source and target PDFs with their discrete empirical distributions supported on mini-batches, making it amenable to implicit distributions and stochastic gradient descent-based optimization. When applied to train a generative model, CT is shown to strike a good balance between mode-covering and mode-seeking behaviors and strongly resist mode collapse. On a wide variety of benchmark datasets for generative modeling, substituting the default statistical distance of an existing generative adversarial network with CT is shown to consistently improve the performance. PyTorch code is provided.
****
Actively Identifying Causal Effects with Latent Variables Given Only Response Va

riable Observable
Tian-Zuo Wang, Zhi-Hua Zhou
In many real tasks, it is generally desired to study the causal effect on a spec ific target (response variable) only, with no need to identify the thorough caus al effects involving all variables. In this paper, we attempt to identify such e ffects by a few active interventions where only the response variable is observa ble. This task is challenging because the causal graph is unknown and even there may exist latent confounders. To learn the necessary structure for identifying the effects, we provide the graphical characterization that allows us to efficie ntly estimate all possible causal effects in a partially mixed ancestral graph (PMAG) by generalized back-door criterion. The characterization guides learning a local structure with the interventional data. Theoretical analysis and empirica l studies validate the effectiveness and efficiency of our proposed approach.
****
Interventional Sum-Product Networks: Causal Inference with Tractable Probabilist ic Models
Matej Ze■evi■, Devendra Dhami, Athresh Karanam, Sriraam Natarajan, Kristian Kers ting
While probabilistic models are an important tool for studying causality, doing s o suffers from the intractability of inference. As a step towards tractable caus al models, we consider the problem of learning interventional distributions usin g sum-product networks (SPNs) that are over-parameterized by gate functions, e.g ., neural networks. Providing an arbitrarily intervened causal graph as input, e ffectively subsuming Pearl's do-operator, the gate function predicts the paramet ers of the SPN. The resulting interventional SPNs are motivated and illustrated by a structural causal model themed around personal health. Our empirical evalua tion against competing methods from both generative and causal modelling demonst rates that interventional SPNs indeed are both expressive and causally adequate.
****
PettingZoo: Gym for Multi-Agent Reinforcement Learning
J Terry, Benjamin Black, Nathaniel Grammel, Mario Jayakumar, Ananth Hari, Ryan Sullivan, Luis S Santos, Clemens Dieffendahl, Caroline Horsch, Rodrigo Perez-Vi cente, Niall Williams, Yashas Lokesh, Praveen Ravi
This paper introduces the PettingZoo library and the accompanying Agent Environm ent Cycle ("AEC") games model. PettingZoo is a library of diverse sets of multi- agent environments with a universal, elegant Python API. PettingZoo was develope d with the goal of accelerating research in Multi-Agent Reinforcement Learning ("MARL"), by making work more interchangeable, accessible and reproducible akin t o what OpenAI's Gym library did for single-agent reinforcement learning. Petting Zoo's API, while inheriting many features of Gym, is unique amongst MARL APIs in that it's based around the novel AEC games model. We argue, in part through cas e studies on major problems in popular MARL environments, that the popular game models are poor conceptual models of the games commonly used with MARL, that the y promote severe bugs that are hard to detect, and that the AEC games model addr esses these problems.
****
Parametric Complexity Bounds for Approximating PDEs with Neural Networks
Tanya Marwah, Zachary Lipton, Andrej Risteski
Requests for name changes in the electronic proceedings will be accepted with 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-learn non-convex piecewise-Lipschitz functions
Maria-Florina F. Balcan, Mikhail Khodak, Dravyansh Sharma, Ameet Talwalkar
We analyze the meta-learning of the initialization and step-size of learning alg orithms for piecewise-Lipschitz functions, a non-convex setting with application s to both machine learning and algorithms. Starting from recent regret bounds fo r the exponential forecaster on losses with dispersed discontinuities, we genera lize them to be initialization-dependent and then use this result to propose a p

ractical meta-learning procedure that learns both the initialization and the ste p-size of the algorithm from multiple online learning tasks. Asymptotically, we guarantee that the average regret across tasks scales with a natural notion of t ask-similarity that measures the amount of overlap between near-optimal regions of different tasks. Finally, we instantiate the method and its guarantee in two important settings: robust meta-learning and multi-task data-driven algorithm de sign.
****
Uncertain Decisions Facilitate Better Preference Learning
Cassidy Laidlaw, Stuart Russell
Existing observational approaches for learning human preferences, such as invers e reinforcement learning, usually make strong assumptions about the observabilit y of the human's environment. However, in reality, people make many important de cisions under uncertainty. To better understand preference learning in these cas es, we study the setting of inverse decision theory (IDT), a previously proposed framework where a human is observed making non-sequential binary decisions unde r uncertainty. In IDT, the human's preferences are conveyed through their loss f unction, which expresses a tradeoff between different types of mistakes. We give the first statistical analysis of IDT, providing conditions necessary to identi fy these preferences and characterizing the sample complexity—the number of deci sions that must be observed to learn the tradeoff the human is making to a desir ed precision. Interestingly, we show that it is actually easier to identify pref erences when the decision problem is more uncertain. Furthermore, uncertain deci sion problems allow us to relax the unrealistic assumption that the human is an optimal decision maker but still identify their exact preferences; we give sampl e complexities in this suboptimal case as well. Our analysis contradicts the int uition that partial observability should make preference learning more difficult . It also provides a first step towards understanding and improving preference l earning methods for uncertain and suboptimal humans.
****
Decision Transformer: Reinforcement Learning via Sequence Modeling
Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, Igor Mordatch
We introduce a framework that abstracts Reinforcement Learning (RL) as a sequenc e modeling problem. This allows us to draw upon the simplicity and scalability o f the Transformer architecture, and associated advances in language modeling suc h as GPT-x and BERT. In particular, we present Decision Transformer, an architec ture that casts the problem of RL as conditional sequence modeling. Unlike prior approaches to RL that fit value functions or compute policy gradients, Decision Transformer simply outputs the optimal actions by leveraging a causally masked Transformer. By conditioning an autoregressive model on the desired return (rewa rd), past states, and actions, our Decision Transformer model can generate futur e actions that achieve the desired return. Despite its simplicity, Decision Tran sformer matches or exceeds the performance of state-of-the-art model-free offlin e RL baselines on Atari, OpenAI Gym, and Key-to-Door tasks.
****
Probability Paths and the Structure of Predictions over Time
Zhiyuan Jerry Lin, Hao Sheng, Sharad Goel
In settings ranging from weather forecasts to political prognostications to fina ncial projections, probability estimates of future binary outcomes often evolve over time. For example, the estimated likelihood of rain on a specific day chang es by the hour as new information becomes available. Given a collection of such probability paths, we introduce a Bayesian framework -- which we call the Gaussi an latent information martingale, or GLIM -- for modeling the structure of dynam ic predictions over time. Suppose, for example, that the likelihood of rain in a week is 50%, and consider two hypothetical scenarios. In the first, one expects the forecast to be equally likely to become either 25% or 75% tomorrow; in the second, one expects the forecast to stay constant for the next several days. A t ime-sensitive decision-maker might select a course of action immediately in the latter scenario, but may postpone their decision in the former, knowing that new

information is imminent. We model these trajectories by assuming predictions update according to a latent process of information flow, which is inferred from historical data. In contrast to general methods for time series analysis, this approach preserves important properties of probability paths such as the martingale structure and appropriate amount of volatility and better quantifies future uncertainties around probability paths. We show that GLIM outperforms three popular baseline methods, producing better estimated posterior probability path distributions measured by three different metrics. By elucidating the dynamic structure of predictions over time, we hope to help individuals make more informed choices.
****
Deep Extended Hazard Models for Survival Analysis
Qixian Zhong, Jonas W. Mueller, Jane-Ling Wang
Unlike standard prediction tasks, survival analysis requires modeling right censored data, which must be treated with care. While deep neural networks excel in traditional supervised learning, it remains unclear how to best utilize these models in survival analysis. A key question asks which data-generating assumptions of traditional survival models should be retained and which should be made more flexible via the function-approximating capabilities of neural networks. Rather than estimating the survival function targeted by most existing methods, we introduce a Deep Extended Hazard (DeepEH) model to provide a flexible and general framework for deep survival analysis. The extended hazard model includes the conventional Cox proportional hazards and accelerated failure time models as special cases, so DeepEH subsumes the popular Deep Cox proportional hazard (DeepSurv) and Deep Accelerated Failure Time (DeepAFT) models. We additionally provide theoretical support for the proposed DeepEH model by establishing consistency and convergence rate of the survival function estimator, which underscore the attractive feature that deep learning is able to detect low-dimensional structure of data in high-dimensional space. Numerical experiments also provide evidence that the proposed methods outperform existing statistical and deep learning approaches to survival analysis.
****
TNASP: A Transformer-based NAS Predictor with a Self-evolution Framework
Shun Lu, Jixiang Li, Jianchao Tan, Sen Yang, Ji Liu
Predictor-based Neural Architecture Search (NAS) continues to be an important topic because it aims to mitigate the time-consuming search procedure of traditional NAS methods. A promising performance predictor determines the quality of final searched models in predictor-based NAS methods. Most existing predictor-based methodologies train model-based predictors under a proxy dataset setting, which may suffer from the accuracy decline and the generalization problem, mainly due to their poor abilities to represent spatial topology information of the graph structure data. Besides the poor encoding for spatial topology information, these works did not take advantage of the temporal information such as historical evaluations during training. Thus, we propose a Transformer-based NAS performance predictor, associated with a Laplacian matrix based positional encoding strategy, which better represents topology information and achieves better performance than previous state-of-the-art methods on NAS-Bench-101, NAS-Bench-201, and DARTS search space. Furthermore, we also propose a self-evolution framework that can fully utilize temporal information as guidance. This framework iteratively involves the evaluations of previously predicted results as constraints into current optimization iteration, thus further improving the performance of our predictor. Such framework is model-agnostic, thus can enhance performance on various backbone structures for the prediction task. Our proposed method helped us rank 2nd among all teams in CVPR 2021 NAS Competition Track 2: Performance Prediction Track.
****
Automorphic Equivalence-aware Graph Neural Network
Fengli Xu, Quanming Yao, Pan Hui, Yong Li
Distinguishing the automorphic equivalence of nodes in a graph plays an essential role in many scientific domains, e.g., computational biologist and social netw

ork analysis. However, existing graph neural networks (GNNs) fail to capture suc
h an important property. To make GNN aware of automorphic equivalence, we first
introduce a localized variant of this concept --- ego-centered automorphic equiv
alence (Ego-AE). Then, we design a novel variant of GNN, i.e., GRAPE, that uses
learnable AE-aware aggregators to explicitly differentiate the Ego-AE of each no
de's neighbors with the aids of various subgraph templates. While the design of
subgraph templates can be hard, we further propose a genetic algorithm to automa
tically search them from graph data. Moreover, we theoretically prove that GRAPE
 is expressive in terms of generating distinct representations for nodes with di
fferent Ego-AE features, which fills in a fundamental gap of existing GNN varian
ts. Finally, we empirically validate our model on eight real-world graph data, i
ncluding social network, e-commerce co-purchase network, and citation network, a
nd show that it consistently outperforms existing GNNs. The source code is publi
c available at https://github.com/tsinghua-fib-lab/GRAPE.
****

Random Shuffling Beats SGD Only After Many Epochs on Ill-Conditioned Problems
Itay Safran, Ohad Shamir
Recently, there has been much interest in studying the convergence rates of with
out-replacement SGD, and proving that it is faster than with-replacement SGD in
the worst case. However, known lower bounds ignore the problem's geometry, inclu
ding its condition number, whereas the upper bounds explicitly depend on it. Per
haps surprisingly, we prove that when the condition number is taken into account
, without-replacement SGD \emph{does not} significantly improve on with-replacem
ent SGD in terms of worst-case bounds, unless the number of epochs (passes over
the data) is larger than the condition number. Since many problems in machine le
arning and other areas are both ill-conditioned and involve large datasets, this
 indicates that without-replacement does not necessarily improve over with-repla
cement sampling for realistic iteration budgets. We show this by providing new l
ower and upper bounds which are tight (up to log factors), for quadratic problem
s with commuting quadratic terms, precisely quantifying the dependence on the pr
oblem parameters.
****

Analytic Study of Families of Spurious Minima in Two-Layer ReLU Neural Networks:
 A Tale of Symmetry II
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-auth
ors prior to requesting a name change in the electronic proceedings.
****

CAM-GAN: Continual Adaptation Modules for Generative Adversarial Networks
Sakshi Varshney, Vinay Kumar Verma, P. K. Srijith, Lawrence Carin, Piyush Rai
We present a continual learning approach for generative adversarial networks (GA
Ns), by designing and leveraging parameter-efficient feature map transformations
. Our approach is based on learning a set of global and task-specific parameters
. The global parameters are fixed across tasks whereas the task-specific paramet
ers act as local adapters for each task, and help in efficiently obtaining task-
specific feature maps. Moreover, we propose an element-wise addition of residual
 bias in the transformed feature space, which further helps stabilize GAN traini
ng in such settings. Our approach also leverages task similarities based on the
Fisher information matrix. Leveraging this knowledge from previous tasks signifi
cantly improves the model performance. In addition, the similarity measure also
helps reduce the parameter growth in continual adaptation and helps to learn a c
ompact model. In contrast to the recent approaches for continually-learned GANs,
 the proposed approach provides a memory-efficient way to perform effective cont
inual data generation. Through extensive experiments on challenging and diverse
datasets, we show that the feature-map-transformation approach outperforms state
-of-the-art methods for continually-learned GANs, with substantially fewer param
eters. The proposed method generates high-quality samples that can also improve
the generative-replay-based continual learning for discriminative tasks.

****
Structured Dropout Variational Inference for Bayesian Neural Networks
Son Nguyen, Duong Nguyen, Khai Nguyen, Khoat Than, Hung Bui, Nhat Ho
Approximate inference in Bayesian deep networks exhibits a dilemma of how to yield high fidelity posterior approximations while maintaining computational efficiency and scalability. We tackle this challenge by introducing a novel variational structured approximation inspired by the Bayesian interpretation of Dropout regularization. Concretely, we focus on the inflexibility of the factorized structure in Dropout posterior and then propose an improved method called Variational Structured Dropout (VSD). VSD employs an orthogonal transformation to learn a structured representation on the variational Gaussian noise with plausible complexity, and consequently induces statistical dependencies in the approximate posterior. Theoretically, VSD successfully addresses the pathologies of previous Variational Dropout methods and thus offers a standard Bayesian justification. We further show that VSD induces an adaptive regularization term with several desirable properties which contribute to better generalization. Finally, we conduct extensive experiments on standard benchmarks to demonstrate the effectiveness of VSD over state-of-the-art variational methods on predictive accuracy, uncertainty estimation, and out-of-distribution detection.
****
Neural Relightable Participating Media Rendering
Quan Zheng, Gurprit Singh, Hans-peter Seidel
Learning neural radiance fields of a scene has recently allowed realistic novel view synthesis of the scene, but they are limited to synthesize images under the original fixed lighting condition. Therefore, they are not flexible for the eagerly desired tasks like relighting, scene editing and scene composition. To tackle this problem, several recent methods propose to disentangle reflectance and illumination from the radiance field. These methods can cope with solid objects with opaque surfaces but participating media are neglected. Also, they take into account only direct illumination or at most one-bounce indirect illumination, thus suffer from energy loss due to ignoring the high-order indirect illumination. We propose to learn neural representations for participating media with a complete simulation of global illumination. We estimate direct illumination via ray tracing and compute indirect illumination with spherical harmonics. Our approach avoids computing the lengthy indirect bounces and does not suffer from energy loss. Our experiments on multiple scenes show that our approach achieves superior visual quality and numerical performance compared to state-of-the-art methods, and it can generalize to deal with solid objects with opaque surfaces as well.
****
Efficient Neural Network Training via Forward and Backward Propagation Sparsification
Xiao Zhou, Weizhong Zhang, Zonghao Chen, SHIZHE DIAO, Tong Zhang
Sparse training is a natural idea to accelerate the training speed of deep neural networks and save the memory usage, especially since large modern neural networks are significantly over-parameterized. However, most of the existing methods cannot achieve this goal in practice because the chain rule based gradient (w.r.t. structure parameters) estimators adopted by previous methods require dense computation at least in the backward propagation step. This paper solves this problem by proposing an efficient sparse training method with completely sparse forward and backward passes. We first formulate the training process as a continuous minimization problem under global sparsity constraint. We then separate the optimization process into two steps, corresponding to weight update and structure parameter update. For the former step, we use the conventional chain rule, which can be sparse via exploiting the sparse structure. For the latter step, instead of using the chain rule based gradient estimators as in existing methods, we propose a variance reduced policy gradient estimator, which only requires two forward passes without backward propagation, thus achieving completely sparse training. We prove that the variance of our gradient estimator is bounded. Extensive experimental results on real-world datasets demonstrate that compared to previous methods, our algorithm is much more effective in accelerating the training pr

ocess, up to an order of magnitude faster.
****
Learning to Ground Multi-Agent Communication with Autoencoders
Toru Lin, Jacob Huh, Christopher Stauffer, Ser Nam Lim, Phillip Isola
Communication requires having a common language, a lingua franca, between agents. This language could emerge via a consensus process, but it may require many generations of trial and error. Alternatively, the lingua franca can be given by the environment, where agents ground their language in representations of the observed world. We demonstrate a simple way to ground language in learned representations, which facilitates decentralized multi-agent communication and coordination. We find that a standard representation learning algorithm -- autoencoding -- is sufficient for arriving at a grounded common language. When agents broadcast these representations, they learn to understand and respond to each other's utterances and achieve surprisingly strong task performance across a variety of multi-agent communication environments.
****
Large-Scale Wasserstein Gradient Flows
Petr Mokrov, Alexander Korotin, Lingxiao Li, Aude Genevay, Justin M. Solomon, Evgeny Burnaev
Wasserstein gradient flows provide a powerful means of understanding and solving many diffusion equations. Specifically, Fokker-Planck equations, which model the diffusion of probability measures, can be understood as gradient descent over entropy functionals in Wasserstein space. This equivalence, introduced by Jordan, Kinderlehrer and Otto, inspired the so-called JKO scheme to approximate these diffusion processes via an implicit discretization of the gradient flow in Wasserstein space. Solving the optimization problem associated with each JKO step, however, presents serious computational challenges. We introduce a scalable method to approximate Wasserstein gradient flows, targeted to machine learning applications. Our approach relies on input-convex neural networks (ICNNs) to discretize the JKO steps, which can be optimized by stochastic gradient descent. Contrarily to previous work, our method does not require domain discretization or particle simulation. As a result, we can sample from the measure at each time step of the diffusion and compute its probability density. We demonstrate the performance of our algorithm by computing diffusions following the Fokker-Planck equation and apply it to unnormalized density sampling as well as nonlinear filtering.
****
Who Leads and Who Follows in Strategic Classification?
Tijana Zrnic, Eric Mazumdar, Shankar Sastry, Michael Jordan
As predictive models are deployed into the real world, they must increasingly contend with strategic behavior. A growing body of work on strategic classification treats this problem as a Stackelberg game: the decision-maker "leads" in the game by deploying a model, and the strategic agents "follow" by playing their best response to the deployed model. Importantly, in this framing, the burden of learning is placed solely on the decision-maker, while the agents’ best responses are implicitly treated as instantaneous. In this work, we argue that the order of play in strategic classification is fundamentally determined by the relative frequencies at which the decision-maker and the agents adapt to each other’s actions. In particular, by generalizing the standard model to allow both players to learn over time, we show that a decision-maker that makes updates faster than the agents can reverse the order of play, meaning that the agents lead and the decision-maker follows. We observe in standard learning settings that such a role reversal can be desirable for both the decision-maker and the strategic agents. Finally, we show that a decision-maker with the freedom to choose their update frequency can induce learning dynamics that converge to Stackelberg equilibria with either order of play.
****
Unadversarial Examples: Designing Objects for Robust Vision
Hadi Salman, Andrew Ilyas, Logan Engstrom, Sai Vemprala, Aleksander Madry, Ashish Kapoor
We study a class of computer vision settings wherein one can modify the design o

f the objects being recognized. We develop a framework that leverages this capab
ility---and deep networks' unusual sensitivity to input perturbations---to desig
n ``robust objects,'' i.e., objects that are explicitly optimized to be confiden
tly classified. Our framework yields improved performance on standard benchmarks
, a simulated robotics environment, and physical-world experiments.
****
Deep Jump Learning for Off-Policy Evaluation in Continuous Treatment Settings
Hengrui Cai, Chengchun Shi, Rui Song, Wenbin Lu
We consider off-policy evaluation (OPE) in continuous treatment settings, such a
s personalized dose-finding. In OPE, one aims to estimate the mean outcome under
 a new treatment decision rule using historical data generated by a different de
cision rule. Most existing works on OPE focus on discrete treatment settings. To
 handle continuous treatments, we develop a novel estimation method for OPE usin
g deep jump learning. The key ingredient of our method lies in adaptively discre
tizing the treatment space using deep discretization, by leveraging deep learnin
g and multi-scale change point detection. This allows us to apply existing OPE m
ethods in discrete treatments to handle continuous treatments. Our method is fur
ther justified by theoretical results, simulations, and a real application to Wa
rfarin Dosing.
****
Attention Approximates Sparse Distributed Memory
Trenton Bricken, Cengiz Pehlevan
While Attention has come to be an important mechanism in deep learning, there re
mains limited intuition for why it works so well. Here, we show that Transformer
 Attention can be closely related under certain data conditions to Kanerva's Spa
rse Distributed Memory (SDM), a biologically plausible associative memory model.
 We confirm that these conditions are satisfied in pre-trained GPT2 Transformer
models. We discuss the implications of the Attention-SDM map and provide new com
putational and biological interpretations of Attention.
****
Augmented Shortcuts for Vision Transformers
Yehui Tang, Kai Han, Chang Xu, An Xiao, Yiping Deng, Chao Xu, Yunhe Wang
Transformer models have achieved great progress on computer vision tasks recentl
y. The rapid development of vision transformers is mainly contributed by their h
igh representation ability for extracting informative features from input images
. However, the mainstream transformer models are designed with deep architecture
s, and the feature diversity will be continuously reduced as the depth increases
, \ie, feature collapse. In this paper, we theoretically analyze the feature col
lapse phenomenon and study the relationship between shortcuts and feature divers
ity in these transformer models. Then, we present an augmented shortcut scheme,
which inserts additional paths with learnable parameters in parallel on the orig
inal shortcuts. To save the computational costs, we further explore an efficient
 approach that uses the block-circulant projection to implement augmented shortc
uts. Extensive experiments conducted on benchmark datasets demonstrate the effec
tiveness of the proposed method, which brings about 1% accuracy increase of the
state-of-the-art visual transformers without obviously increasing their paramete
rs and FLOPs.
****
Finding Regions of Heterogeneity in Decision-Making via Expected Conditional Cov
ariance
Justin Lim, Christina X Ji, Michael Oberst, Saul Blecker, Leora Horwitz, David S
ontag
Individuals often make different decisions when faced with the same context, due
 to personal preferences and background. For instance, judges may vary in their
 leniency towards certain drug-related offenses, and doctors may vary in their p
reference for how to start treatment for certain types of patients. With these
examples in mind, we present an algorithm for identifying types of contexts (e.g
., types of cases or patients) with high inter-decision-maker disagreement. We
formalize this as a causal inference problem, seeking a region where the assignm
ent of decision-maker has a large causal effect on the decision. Our algorithm

finds such a region by maximizing an empirical objective, and we give a generalization bound for its performance. In a semi-synthetic experiment, we show that our algorithm recovers the correct region of heterogeneity accurately compared to baselines. Finally, we apply our algorithm to real-world healthcare datasets, recovering variation that aligns with existing clinical knowledge.
****
Identifying and Benchmarking Natural Out-of-Context Prediction Problems
David Madras, Richard Zemel
Deep learning systems frequently fail at out-of-context (OOC) prediction, the problem of making reliable predictions on uncommon or unusual inputs or subgroups of the training distribution. To this end, a number of benchmarks for measuring OOC performance have been recently introduced. In this work, we introduce a framework unifying the literature on OOC performance measurement, and demonstrate how rich auxiliary information can be leveraged to identify candidate sets of OOC examples in existing datasets. We present NOOCh: a suite of naturally-occurring "challenge sets", and show how varying notions of context can be used to probe specific OOC failure modes. Experimentally, we explore the tradeoffs between various learning approaches on these challenge sets and demonstrate how the choices made in designing OOC benchmarks can yield varying conclusions.
****
Label Disentanglement in Partition-based Extreme Multilabel Classification
Xuanqing Liu, Wei-Cheng Chang, Hsiang-Fu Yu, Cho-Jui Hsieh, Inderjit Dhillon
Partition-based methods are increasingly-used in extreme multi-label classification (XMC) problems due to their scalability to large output spaces (e.g., millions or more). However, existing methods partition the large label space into mutually exclusive clusters, which is sub-optimal when labels have multi-modality and rich semantics. For instance, the label “Apple” can be the fruit or the brand name, which leads to the following research question: can we disentangle these multi-modal labels with non-exclusive clustering tailored for downstream XMC tasks? In this paper, we show that the label assignment problem in partition-based XMC can be formulated as an optimization problem, with the objective of maximizing precision rates. This leads to an efficient algorithm to form flexible and overlapped label clusters, and a method that can alternatively optimizes the cluster assignments and the model parameters for partition-based XMC. Experimental results on synthetic and real datasets show that our method can successfully disentangle multi-modal labels, leading to state-of-the-art (SOTA) results on four XMC benchmarks.
****
Leveraging SE(3) Equivariance for Self-supervised Category-Level Object Pose Estimation from Point Clouds
Xiaolong Li, Yijia Weng, Li Yi, Leonidas J. Guibas, A. Abbott, Shuran Song, He Wang
Category-level object pose estimation aims to find 6D object poses of previously unseen object instances from known categories without access to object CAD models. To reduce the huge amount of pose annotations needed for category-level learning, we propose for the first time a self-supervised learning framework to estimate category-level 6D object pose from single 3D point clouds. During training, our method assumes no ground-truth pose annotations, no CAD models, and no multi-view supervision. The key to our method is to disentangle shape and pose through an invariant shape reconstruction module and an equivariant pose estimation module, empowered by SE(3) equivariant point cloud networks. The invariant shape reconstruction module learns to perform aligned reconstructions, yielding a category-level reference frame without using any annotations. In addition, the equivariant pose estimation module achieves category-level pose estimation accuracy that is comparable to some fully supervised methods. Extensive experiments demonstrate the effectiveness of our approach on both complete and partial depth point clouds from the ModelNet40 benchmark, and on real depth point clouds from the NOCS-REAL 275 dataset. The project page with code and visualizations can be found at: dragonlong.github.io/equi-pose.
****

A Theoretical Analysis of Fine-tuning with Linear Teachers
Gal Shachaf, Alon Brutzkus, Amir Globerson
Fine-tuning is a common practice in deep learning, achieving excellent generalization results on downstream tasks using relatively little training data. Although widely used in practice, it is not well understood theoretically. Here we analyze the sample complexity of this scheme for regression with linear teachers in several settings. Intuitively, the success of fine-tuning depends on the similarity between the source tasks and the target task. But what is the right way of measuring this similarity? We show that the relevant measure has to do with the relation between the source task, the target task and the covariance structure of the target data. In the setting of linear regression, we show that under realistic settings there can be substantial sample complexity reduction when the above measure is low. For deep linear regression, we propose a novel result regarding the inductive bias of gradient-based training when the network is initialized with pretrained weights. Using this result we show that the similarity measure for this setting is also affected by the depth of the network. We conclude with results on shallow ReLU models, and analyze the dependence of sample complexity there on source and target tasks. We empirically demonstrate our results for both synthetic and realistic data.
****
Overinterpretation reveals image classification model pathologies
Brandon Carter, Siddhartha Jain, Jonas W. Mueller, David Gifford
Image classifiers are typically scored on their test set accuracy, but high accuracy can mask a subtle type of model failure. We find that high scoring convolutional neural networks (CNNs) on popular benchmarks exhibit troubling pathologies that allow them to display high accuracy even in the absence of semantically salient features. When a model provides a high-confidence decision without salient supporting input features, we say the classifier has overinterpreted its input, finding too much class-evidence in patterns that appear nonsensical to humans. Here, we demonstrate that neural networks trained on CIFAR-10 and ImageNet suffer from overinterpretation, and we find models on CIFAR-10 make confident predictions even when 95% of input images are masked and humans cannot discern salient features in the remaining pixel-subsets. We introduce Batched Gradient SIS, a new method for discovering sufficient input subsets for complex datasets, and use this method to show the sufficiency of border pixels in ImageNet for training and testing. Although these patterns portend potential model fragility in real-world deployment, they are in fact valid statistical patterns of the benchmark that alone suffice to attain high test accuracy. Unlike adversarial examples, overinterpretation relies upon unmodified image pixels. We find ensembling and input dropout can each help mitigate overinterpretation.
****
Neural Circuit Synthesis from Specification Patterns
Frederik Schmitt, Christopher Hahn, Markus N Rabe, Bernd Finkbeiner
We train hierarchical Transformers on the task of synthesizing hardware circuits directly out of high-level logical speci■cations in linear-time temporal logic (LTL). The LTL synthesis problem is a well-known algorithmic challenge with a long history and an annual competition is organized to track the improvement of algorithms and tooling over time. New approaches using machine learning might open a lot of possibilities in this area, but suffer from the lack of suf■cient amounts of training data. In this paper, we consider a method to generate large amounts of additional training data, i.e., pairs of speci■cations and circuits implementing them. We ensure that this synthetic data is suf■ciently close to human-written speci■cations by mining common patterns from the speci■cations used in the synthesis competitions. We show that hierarchical Transformers trained on this synthetic data solve a signi■cant portion of problems from the synthesis competitions, and even out-of-distribution examples from a recent case study.
****
Directional Message Passing on Molecular Graphs via Synthetic Coordinates
Johannes Gasteiger, Chandan Yeshwanth, Stephan Günnemann
Graph neural networks that leverage coordinates via directional message passing

have recently set the state of the art on multiple molecular property prediction tasks. However, they rely on atom position information that is often unavailable, and obtaining it is usually prohibitively expensive or even impossible. In this paper we propose synthetic coordinates that enable the use of advanced GNNs without requiring the true molecular configuration. We propose two distances as synthetic coordinates: Distance bounds that specify the rough range of molecular configurations, and graph-based distances using a symmetric variant of personalized PageRank. To leverage both distance and angular information we propose a method of transforming normal graph neural networks into directional MPNNs. We show that with this transformation we can reduce the error of a normal graph neural network by 55% on the ZINC benchmark. We furthermore set the state of the art on ZINC and coordinate-free QM9 by incorporating synthetic coordinates in the SMP and DimeNet++ models. Our implementation is available online.
****

Federated Multi-Task Learning under a Mixture of Distributions
Othmane Marfoq, Giovanni Neglia, Aurélien Bellet, Laetitia Kameni, Richard Vidal
The increasing size of data generated by smartphones and IoT devices motivated the development of Federated Learning (FL), a framework for on-device collaborative training of machine learning models. First efforts in FL focused on learning a single global model with good average performance across clients, but the global model may be arbitrarily bad for a given client, due to the inherent heterogeneity of local data distributions. Federated multi-task learning (MTL) approaches can learn personalized models by formulating an opportune penalized optimization problem. The penalization term can capture complex relations among personalized models, but eschews clear statistical assumptions about local data distributions. In this work, we propose to study federated MTL under the flexible assumption that each local data distribution is a mixture of unknown underlying distributions. This assumption encompasses most of the existing personalized FL approaches and leads to federated EM-like algorithms for both client-server and fully decentralized settings. Moreover, it provides a principled way to serve personalized models to clients not seen at training time. The algorithms' convergence is analyzed through a novel federated surrogate optimization framework, which can be of general interest. Experimental results on FL benchmarks show that our approach provides models with higher accuracy and fairness than state-of-the-art methods.
****

Learning Generative Vision Transformer with Energy-Based Latent Space for Saliency Prediction
Jing Zhang, Jianwen Xie, Nick Barnes, Ping Li
Vision transformer networks have shown superiority in many computer vision tasks. In this paper, we take a step further by proposing a novel generative vision transformer with latent variables following an informative energy-based prior for salient object detection. Both the vision transformer network and the energy-based prior model are jointly trained via Markov chain Monte Carlo-based maximum likelihood estimation, in which the sampling from the intractable posterior and prior distributions of the latent variables are performed by Langevin dynamics. Further, with the generative vision transformer, we can easily obtain a pixel-wise uncertainty map from an image, which indicates the model confidence in predicting saliency from the image. Different from the existing generative models which define the prior distribution of the latent variables as a simple isotropic Gaussian distribution, our model uses an energy-based informative prior which can be more expressive to capture the latent space of the data. We apply the proposed framework to both RGB and RGB-D salient object detection tasks. Extensive experimental results show that our framework can achieve not only accurate saliency predictions but also meaningful uncertainty maps that are consistent with the human perception.
****

Regularization in ResNet with Stochastic Depth
Soufiane Hayou, Fadhel Ayed
Regularization plays a major role in modern deep learning. From classic techniqu

es such as L1, L2 penalties to other noise-based methods such as Dropout, regularization often yields better generalization properties by avoiding overfitting. Recently, Stochastic Depth (SD) has emerged as an alternative regularization technique for residual neural networks (ResNets) and has proven to boost the performance of ResNet on many tasks [Huang et al., 2016]. Despite the recent success of SD, little is known about this technique from a theoretical perspective. This paper provides a hybrid analysis combining perturbation analysis and signal propagation to shed light on different regularization effects of SD. Our analysis allows us to derive principled guidelines for choosing the survival rates used for training with SD.
****
ResT: An Efficient Transformer for Visual Recognition
Qinglong Zhang, Yu-Bin Yang
This paper presents an efficient multi-scale vision Transformer, called ResT, that capably served as a general-purpose backbone for image recognition. Unlike existing Transformer methods, which employ standard Transformer blocks to tackle raw images with a fixed resolution, our ResT have several advantages: (1) A memory-efficient multi-head self-attention is built, which compresses the memory by a simple depth-wise convolution, and projects the interaction across the attention-heads dimension while keeping the diversity ability of multi-heads; (2) Positional encoding is constructed as spatial attention, which is more flexible and can tackle with input images of arbitrary size without interpolation or fine-tune; (3) Instead of the straightforward tokenization at the beginning of each stage, we design the patch embedding as a stack of overlapping convolution operation with stride on the token map. We comprehensively validate ResT on image classification and downstream tasks. Experimental results show that the proposed ResT can outperform the recently state-of-the-art backbones by a large margin, demonstrating the potential of ResT as strong backbones. The code and models will be made publicly available at https://github.com/wofmanaf/ResT.
****
Adversarial Examples for k-Nearest Neighbor Classifiers Based on Higher-Order Voronoi Diagrams
Chawin Sitawarin, Evgenios Kornaropoulos, Dawn Song, David Wagner
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 Robust 3D Point Cloud Recognition Using Self-Supervisions
Jiachen Sun, Yulong Cao, Christopher B Choy, Zhiding Yu, Anima Anandkumar, Zhuoqing Morley Mao, Chaowei Xiao
3D point cloud data is increasingly used in safety-critical applications such as autonomous driving. Thus, the robustness of 3D deep learning models against adversarial attacks becomes a major consideration. In this paper, we systematically study the impact of various self-supervised learning proxy tasks on different architectures and threat models for 3D point clouds with adversarial training. Specifically, we study MLP-based (PointNet), convolution-based (DGCNN), and transformer-based (PCT) 3D architectures. Through extensive experimentation, we demonstrate that appropriate applications of self-supervision can significantly enhance the robustness in 3D point cloud recognition, achieving considerable improvements compared to the standard adversarial training baseline. Our analysis reveals that local feature learning is desirable for adversarial robustness in point clouds since it limits the adversarial propagation between the point-level input perturbations and the model's final output. This insight also explains the success of DGCNN and the jigsaw proxy task in achieving stronger 3D adversarial robustness.
****
Tuning Mixed Input Hyperparameters on the Fly for Efficient Population Based AutoRL
Jack Parker-Holder, Vu Nguyen, Shaan Desai, Stephen J Roberts

Despite a series of recent successes in reinforcement learning (RL), many RL algorithms remain sensitive to hyperparameters. As such, there has recently been interest in the field of AutoRL, which seeks to automate design decisions to create more general algorithms. Recent work suggests that population based approaches may be effective AutoRL algorithms, by learning hyperparameter schedules on the fly. In particular, the PB2 algorithm is able to achieve strong performance in RL tasks by formulating online hyperparameter optimization as time varying GP-bandit problem, while also providing theoretical guarantees. However, PB2 is only designed to work for \emph{continuous} hyperparameters, which severely limits its utility in practice. In this paper we introduce a new (provably) efficient hierarchical approach for optimizing \emph{both continuous and categorical} variables, using a new time-varying bandit algorithm specifically designed for the population based training regime. We evaluate our approach on the challenging Procgen benchmark, where we show that explicitly modelling dependence between data augmentation and other hyperparameters improves generalization.
****
Neural Algorithmic Reasoners are Implicit Planners
Andreea-Ioana Deac, Petar Veli■kovi■, Ognjen Milinkovic, Pierre-Luc Bacon, Jian Tang, Mladen Nikolic
Implicit planning has emerged as an elegant technique for combining learned models of the world with end-to-end model-free reinforcement learning. We study the class of implicit planners inspired by value iteration, an algorithm that is guaranteed to yield perfect policies in fully-specified tabular environments. We find that prior approaches either assume that the environment is provided in such a tabular form---which is highly restrictive---or infer "local neighbourhoods" of states to run value iteration over---for which we discover an algorithmic bottleneck effect. This effect is caused by explicitly running the planning algorithm based on scalar predictions in every state, which can be harmful to data efficiency if such scalars are improperly predicted. We propose eXecuted Latent Value Iteration Networks (XLVINs), which alleviate the above limitations. Our method performs all planning computations in a high-dimensional latent space, breaking the algorithmic bottleneck. It maintains alignment with value iteration by carefully leveraging neural graph-algorithmic reasoning and contrastive self-supervised learning. Across seven low-data settings---including classical control, navigation and Atari---XLVINs provide significant improvements to data efficiency against value iteration-based implicit planners, as well as relevant model-free baselines. Lastly, we empirically verify that XLVINs can closely align with value iteration.
****
Self-Supervised Learning with Kernel Dependence Maximization
Yazhe Li, Roman Pogodin, Danica J. Sutherland, Arthur Gretton
We approach self-supervised learning of image representations from a statistical dependence perspective, proposing Self-Supervised Learning with the Hilbert-Schmidt Independence Criterion (SSL-HSIC). SSL-HSIC maximizes dependence between representations of transformations of an image and the image identity, while minimizing the kernelized variance of those representations. This framework yields a new understanding of InfoNCE, a variational lower bound on the mutual information (MI) between different transformations. While the MI itself is known to have pathologies which can result in learning meaningless representations, its bound is much better behaved: we show that it implicitly approximates SSL-HSIC (with a slightly different regularizer).Our approach also gives us insight into BYOL, a negative-free SSL method, since SSL-HSIC similarly learns local neighborhoods of samples. SSL-HSIC allows us to directly optimize statistical dependence in time linear in the batch size, without restrictive data assumptions or indirect mutual information estimators. Trained with or without a target network, SSL-HSIC matches the current state-of-the-art for standard linear evaluation on ImageNet, semi-supervised learning and transfer to other classification and vision tasks such as semantic segmentation, depth estimation and object recognition. Code is available at https://github.com/deepmind/ssl_hsic.
****

CROCS: Clustering and Retrieval of Cardiac Signals Based on Patient Disease Clas
s, Sex, and Age
Dani Kiyasseh, Tingting Zhu, David Clifton
The process of manually searching for relevant instances in, and extracting info
rmation from, clinical databases underpin a multitude of clinical tasks. Such ta
sks include disease diagnosis, clinical trial recruitment, and continuing medica
l education. This manual search-and-extract process, however, has been hampered
by the growth of large-scale clinical databases and the increased prevalence of
unlabelled instances. To address this challenge, we propose a supervised contras
tive learning framework, CROCS, where representations of cardiac signals associa
ted with a set of patient-specific attributes (e.g., disease class, sex, age) ar
e attracted to learnable embeddings entitled clinical prototypes. We exploit suc
h prototypes for both the clustering and retrieval of unlabelled cardiac signals
 based on multiple patient attributes. We show that CROCS outperforms the state-
of-the-art method, DTC, when clustering and also retrieves relevant cardiac sign
als from a large database. We also show that clinical prototypes adopt a semanti
cally meaningful arrangement based on patient attributes and thus confer a high
degree of interpretability.
****
Representing Hyperbolic Space Accurately using Multi-Component Floats
Tao Yu, Christopher M. De Sa
Hyperbolic space is particularly useful for embedding data with hierarchical str
ucture; however, representing hyperbolic space with ordinary floating-point numb
ers greatly affects the performance due to its \emph{ineluctable} numerical erro
rs. Simply increasing the precision of floats fails to solve the problem and inc
urs a high computation cost for simulating greater-than-double-precision floats
on hardware such as GPUs, which does not support them. In this paper, we propose
 a simple, feasible-on-GPUs, and easy-to-understand solution for numerically acc
urate learning on hyperbolic space. We do this with a new approach to represent
hyperbolic space using multi-component floating-point (MCF) in the Poincar{\'e}
upper-half space model. Theoretically and experimentally we show our model has s
mall numerical error, and on embedding tasks across various datasets, models rep
resented by multi-component floating-points gain more capacity and run significa
ntly faster on GPUs than prior work.
****
Dimensionality Reduction for Wasserstein Barycenter
Zachary Izzo, Sandeep Silwal, Samson 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.
****
Neural Population Geometry Reveals the Role of Stochasticity in Robust Perceptio
n
Joel Dapello, Jenelle Feather, Hang Le, Tiago Marques, David Cox, Josh McDermott
, James J DiCarlo, Sueyeon Chung
Adversarial examples are often cited by neuroscientists and machine learning res
earchers as an example of how computational models diverge from biological senso
ry systems. Recent work has proposed adding biologically-inspired components to
visual neural networks as a way to improve their adversarial robustness. One sur
prisingly effective component for reducing adversarial vulnerability is response
 stochasticity, like that exhibited by biological neurons. Here, using recently
developed geometrical techniques from computational neuroscience, we investigate
 how adversarial perturbations influence the internal representations of standar
d, adversarially trained, and biologically-inspired stochastic networks. We find
 distinct geometric signatures for each type of network, revealing different mec
hanisms for achieving robust representations. Next, we generalize these results
to the auditory domain, showing that neural stochasticity also makes auditory mo
dels more robust to adversarial perturbations. Geometric analysis of the stochas
tic networks reveals overlap between representations of clean and adversarially

perturbed stimuli, and quantitatively demonstrate that competing geometric effects of stochasticity mediate a tradeoff between adversarial and clean performance. Our results shed light on the strategies of robust perception utilized by adversarially trained and stochastic networks, and help explain how stochasticity may be beneficial to machine and biological computation.
****

Unsupervised Learning of Compositional Energy Concepts

Yilun Du, Shuang Li, Yash Sharma, Josh Tenenbaum, Igor Mordatch

Humans are able to rapidly understand scenes by utilizing concepts extracted from prior experience. Such concepts are diverse, and include global scene descriptors, such as the weather or lighting, as well as local scene descriptors, such as the color or size of a particular object. So far, unsupervised discovery of concepts has focused on either modeling the global scene-level or the local object-level factors of variation, but not both. In this work, we propose COMET, which discovers and represents concepts as separate energy functions, enabling us to represent both global concepts as well as objects under a unified framework. COMET discovers energy functions through recomposing the input image, which we find captures independent factors without additional supervision. Sample generation in COMET is formulated as an optimization process on underlying energy functions, enabling us to generate images with permuted and composed concepts. Finally, discovered visual concepts in COMET generalize well, enabling us to compose concepts between separate modalities of images as well as with other concepts discovered by a separate instance of COMET trained on a different dataset. Code and data available at https://energy-based-model.github.io/comet/.
****

Nearly Horizon-Free Offline Reinforcement Learning

Tongzheng Ren, Jialian Li, Bo Dai, Simon S. Du, Sujay Sanghavi

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

Combinatorial Optimization for Panoptic Segmentation: A Fully Differentiable Approach

Ahmed Abbas, Paul Swoboda

We propose a fully differentiable architecture for simultaneous semantic and instance segmentation (a.k.a. panoptic segmentation) consisting of a convolutional neural network and an asymmetric multiway cut problem solver. The latter solves a combinatorial optimization problem that elegantly incorporates semantic and boundary predictions to produce a panoptic labeling. Our formulation allows to directly maximize a smooth surrogate of the panoptic quality metric by backpropagating the gradient through the optimization problem. Experimental evaluation shows improvement by backpropagating through the optimization problem w.r.t. comparable approaches on Cityscapes and COCO datasets. Overall, our approach of combinatorial optimization for panoptic segmentation (COPS) shows the utility of using optimization in tandem with deep learning in a challenging large scale real-world problem and showcases benefits and insights into training such an architecture.
****

Reinforcement Learning with State Observation Costs in Action-Contingent Noiselessly Observable Markov Decision Processes

HyunJi Alex Nam, Scott Fleming, Emma Brunskill

Many real-world problems that require making optimal sequences of decisions under uncertainty involve costs when the agent wishes to obtain information about its environment. We design and analyze algorithms for reinforcement learning (RL) in Action-Contingent Noiselessly Observable MDPs (ACNO-MDPs), a special class of POMDPs in which the agent can choose to either (1) fully observe the state at a cost and then act; or (2) act without any immediate observation information, relying on past observations to infer the underlying state. ACNO-MDPs arise frequently in important real-world application domains like healthcare, in which clinicians must balance the value of information gleaned from medical tests (e.g., bl

ood-based biomarkers) with the costs of gathering that information (e.g., the co sts of labor and materials required to administer such tests). We develop a PAC RL algorithm for tabular ACNO-MDPs that provides substantially tighter bounds, c ompared to generic POMDP-RL algorithms, on the total number of episodes exhibiti ng worse than near-optimal performance. For continuous-state ACNO-MDPs, we propo se a novel method of incorporating observation information that, when coupled wi th modern RL algorithms, yields significantly faster learning compared to other POMDP-RL algorithms in several simulated environments.
****

Iterative Amortized Policy Optimization
Joseph Marino, Alexandre Piche, Alessandro Davide Ialongo, Yisong Yue
Policy networks are a central feature of deep reinforcement learning (RL) algori thms for continuous control, enabling the estimation and sampling of high-value actions. From the variational inference perspective on RL, policy networks, when used with entropy or KL regularization, are a form of amortized optimization, o ptimizing network parameters rather than the policy distributions directly. Howe ver, direct amortized mappings can yield suboptimal policy estimates and restric ted distributions, limiting performance and exploration. Given this perspective, we consider the more flexible class of iterative amortized optimizers. We demon strate that the resulting technique, iterative amortized policy optimization, yi elds performance improvements over direct amortization on benchmark continuous c ontrol tasks.
****

Revisiting the Calibration of Modern Neural Networks
Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, Mario Lucic
Accurate estimation of predictive uncertainty (model calibration) is essential f or the safe application of neural networks. Many instances of miscalibration in modern neural networks have been reported, suggesting a trend that newer, more a ccurate models produce poorly calibrated predictions. Here, we revisit this ques tion for recent state-of-the-art image classification models. We systematically relate model calibration and accuracy, and find that the most recent models, not ably those not using convolutions, are among the best calibrated. Trends observe d in prior model generations, such as decay of calibration with distribution shi ft or model size, are less pronounced in recent architectures. We also show that model size and amount of pretraining do not fully explain these differences, su ggesting that architecture is a major determinant of calibration properties.
****

The decomposition of the higher-order homology embedding constructed from the k-Laplacian
Yu-Chia Chen, Marina Meila
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-auth ors prior to requesting a name change in the electronic proceedings.
****

Breaking the Moments Condition Barrier: No-Regret Algorithm for Bandits with Sup er Heavy-Tailed Payoffs
Han Zhong, Jiayi Huang, Lin Yang, Liwei Wang
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-auth ors prior to requesting a name change in the electronic proceedings.
****

A nonparametric method for gradual change problems with statistical guarantees
Lizhen Nie, Dan Nicolae
We consider the detection and localization of gradual changes in the distributio n of a sequence of time-ordered observations. Existing literature focuses mostly on the simpler abrupt setting which assumes a discontinuity jump in distributio n, and is unrealistic for some applied settings. We propose a general method for

detecting and localizing gradual changes that does not require any specific dat
a generating model, any particular data type, or any prior knowledge about which
 features of the distribution are subject to change. Despite relaxed assumptions
, the proposed method possesses proven theoretical guarantees for both detection
 and localization.
****
Nested Graph Neural Networks
Muhan Zhang, Pan Li
Graph neural network (GNN)'s success in graph classification is closely related
to the Weisfeiler-Lehman (1-WL) algorithm. By iteratively aggregating neighborin
g node features to a center node, both 1-WL and GNN obtain a node representation
 that encodes a rooted subtree around the center node. These rooted subtree repr
esentations are then pooled into a single representation to represent the whole
graph. However, rooted subtrees are of limited expressiveness to represent a non
-tree graph. To address it, we propose Nested Graph Neural Networks (NGNNs). NGN
N represents a graph with rooted subgraphs instead of rooted subtrees, so that t
wo graphs sharing many identical subgraphs (rather than subtrees) tend to have s
imilar representations. The key is to make each node representation encode a sub
graph around it more than a subtree. To achieve this, NGNN extracts a local subg
raph around each node and applies a base GNN to each subgraph to learn a subgrap
h representation. The whole-graph representation is then obtained by pooling the
se subgraph representations. We provide a rigorous theoretical analysis showing
that NGNN is strictly more powerful than 1-WL. In particular, we proved that NGN
N can discriminate almost all r-regular graphs, where 1-WL always fails. Moreove
r, unlike other more powerful GNNs, NGNN only introduces a constant-factor highe
r time complexity than standard GNNs. NGNN is a plug-and-play framework that can
 be combined with various base GNNs. We test NGNN with different base GNNs on se
veral benchmark datasets. NGNN uniformly improves their performance and shows hi
ghly competitive performance on all datasets.
****
Multimodal and Multilingual Embeddings for Large-Scale Speech Mining
Paul-Ambroise Duquenne, Hongyu Gong, Holger Schwenk
We present an approach to encode a speech signal into a fixed-size representatio
n which minimizes the cosine loss with the existing massively multilingual LASER
 text embedding space. Sentences are close in this embedding space, independentl
y of their language and modality, either text or audio. Using a similarity metri
c in that multimodal embedding space, we perform mining of audio in German, Fren
ch, Spanish and English from Librivox against billions of sentences from Common
Crawl. This yielded more than twenty thousand hours of aligned speech translatio
ns. To evaluate the automatically mined speech/text corpora, we train neural sp
eech translation systems for several languages pairs. Adding the mined data, ach
ieves significant improvements in the BLEU score on the CoVoST2 and the MUST-C t
est sets with respect to a very competitive baseline. Our approach can also be u
sed to directly perform speech-to-speech mining, without the need to first trans
cribe or translate the data. We obtain more than one thousand three hundred hour
s of aligned speech in French, German, Spanish and English. This speech corpus h
as the potential to boost research in speech-to-speech translation which suffers
 from scarcity of natural end-to-end training data. All the mined multimodal cor
pora will be made freely available.
****
Necessary and sufficient graphical conditions for optimal adjustment sets in cau
sal graphical models with hidden variables
Jakob Runge
The problem of selecting optimal backdoor adjustment sets to estimate causal eff
ects in graphical models with hidden and conditioned variables is addressed. Pre
vious work has defined optimality as achieving the smallest asymptotic estimatio
n variance and derived an optimal set for the case without hidden variables. For
 the case with hidden variables there can be settings where no optimal set exist
s and currently only a sufficient graphical optimality criterion of limited appl
icability has been derived. In the present work optimality is characterized as m

aximizing a certain adjustment information which allows to derive a necessary and sufficient graphical criterion for the existence of an optimal adjustment set and a definition and algorithm to construct it. Further, the optimal set is valid if and only if a valid adjustment set exists and has higher (or equal) adjustment information than the Adjust-set proposed in Perkovi{\'c} et~al. [Journal of Machine Learning Research, 18: 1--62, 2018] for any graph. The results translate to minimal asymptotic estimation variance for a class of estimators whose asymptotic variance follows a certain information-theoretic relation. Numerical experiments indicate that the asymptotic results also hold for relatively small sample sizes and that the optimal adjustment set or minimized variants thereof often yield better variance also beyond that estimator class. Surprisingly, among the randomly created setups more than 90\% fulfill the optimality conditions indicating that also in many real-world scenarios graphical optimality may hold.
****
On Blame Attribution for Accountable Multi-Agent Sequential Decision Making
Stelios Triantafyllou, Adish Singla, Goran Radanovic
Blame attribution is one of the key aspects of accountable decision making, as it provides means to quantify the responsibility of an agent for a decision making outcome. In this paper, we study blame attribution in the context of cooperative multi-agent sequential decision making. As a particular setting of interest, we focus on cooperative decision making formalized by Multi-Agent Markov Decision Processes (MMDPs), and we analyze different blame attribution methods derived from or inspired by existing concepts in cooperative game theory. We formalize desirable properties of blame attribution in the setting of interest, and we analyze the relationship between these properties and the studied blame attribution methods. Interestingly, we show that some of the well known blame attribution methods, such as Shapley value, are not performance-incentivizing, while others, such as Banzhaf index, may over-blame agents. To mitigate these value misalignment and fairness issues, we introduce a novel blame attribution method, unique in the set of properties it satisfies, which trade-offs explanatory power (by under-blaming agents) for the aforementioned properties. We further show how to account for uncertainty about agents' decision making policies, and we experimentally: a) validate the qualitative properties of the studied blame attribution methods, and b) analyze their robustness to uncertainty.
****
FLEX: Unifying Evaluation for Few-Shot NLP
Jonathan Bragg, Arman Cohan, Kyle Lo, Iz Beltagy
Few-shot NLP research is highly active, yet conducted in disjoint research threads with evaluation suites that lack challenging-yet-realistic testing setups and fail to employ careful experimental design. Consequently, the community does not know which techniques perform best or even if they outperform simple baselines. In response, we formulate the FLEX Principles, a set of requirements and best practices for unified, rigorous, valid, and cost-sensitive few-shot NLP evaluation. These principles include Sample Size Design, a novel approach to benchmark design that optimizes statistical accuracy and precision while keeping evaluation costs manageable. Following the principles, we release the FLEX benchmark, which includes four few-shot transfer settings, zero-shot evaluation, and a public leaderboard that covers diverse NLP tasks. In addition, we present UniFew, a prompt-based model for few-shot learning that unifies pretraining and finetuning prompt formats, eschewing complex machinery of recent prompt-based approaches in adapting downstream task formats to language model pretraining objectives. We demonstrate that despite simplicity, UniFew achieves results competitive with both popular meta-learning and prompt-based approaches.
****
A flow-based latent state generative model of neural population responses to natural images
Mohammad Bashiri, Edgar Walker, Konstantin-Klemens Lurz, Akshay Jagadish, Taliah Muhammad, Zhiwei Ding, Zhuokun Ding, Andreas Tolias, Fabian Sinz
We present a joint deep neural system identification model for two major sources of neural variability: stimulus-driven and stimulus-conditioned fluctuations. T

o this end, we combine (1) state-of-the-art deep networks for stimulus-driven ac
tivity and (2) a flexible, normalizing flow-based generative model to capture th
e stimulus-conditioned variability including noise correlations. This allows us
to train the model end-to-end without the need for sophisticated probabilistic a
pproximations associated with many latent state models for stimulus-conditioned
fluctuations. We train the model on the responses of thousands of neurons from m
ultiple areas of the mouse visual cortex to natural images. We show that our mod
el outperforms previous state-of-the-art models in predicting the distribution o
f neural population responses to novel stimuli, including shared stimulus-condit
ioned variability. Furthermore, it successfully learns known latent factors of t
he population responses that are related to behavioral variables such as pupil d
ilation, and other factors that vary systematically with brain area or retinotop
ic location. Overall, our model accurately accounts for two critical sources of
neural variability while avoiding several complexities associated with many exis
ting latent state models. It thus provides a useful tool for uncovering the inte
rplay between different factors that contribute to variability in neural activit
y.
****
Learnable Fourier Features for Multi-dimensional Spatial Positional Encoding
Yang Li, Si Si, Gang Li, Cho-Jui Hsieh, Samy Bengio
Requests for name changes in the electronic proceedings will be accepted with no
 questions asked. However name changes may cause bibliographic tracking issues.
  Authors are asked to consider this carefully and discuss it with their co-auth
ors prior to requesting a name change in the electronic proceedings.
****
Doubly Robust Thompson Sampling with Linear Payoffs
Wonyoung Kim, Gi-Soo Kim, Myunghee Cho Paik
Requests for name changes in the electronic proceedings will be accepted with no
 questions asked. However name changes may cause bibliographic tracking issues.
  Authors are asked to consider this carefully and discuss it with their co-auth
ors prior to requesting a name change in the electronic proceedings.
****
A Computationally Efficient Method for Learning Exponential Family Distributions
Abhin Shah, Devavrat Shah, Gregory Wornell
Requests for name changes in the electronic proceedings will be accepted with 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.
****
Rethinking Neural Operations for Diverse Tasks
Nicholas Roberts, Mikhail Khodak, Tri Dao, Liam Li, Christopher Ré, Ameet Talwal
kar
An important goal of AutoML is to automate-away the design of neural networks on
 new tasks in under-explored domains. Motivated by this goal, we study the probl
em of enabling users to discover the right neural operations given data from the
ir specific domain. We introduce a search space of operations called XD-Operatio
ns that mimic the inductive bias of standard multi-channel convolutions while be
ing much more expressive: we prove that it includes many named operations across
 multiple application areas. Starting with any standard backbone such as ResNet,
 we show how to transform it into a search space over XD-operations and how to t
raverse the space using a simple weight sharing scheme. On a diverse set of task
s—solving PDEs, distance prediction for protein folding, and music modeling—our
approach consistently yields models with lower error than baseline networks and
often even lower error than expert-designed domain-specific approaches.
****
Motif-based Graph Self-Supervised Learning for Molecular Property Prediction
ZAIXI ZHANG, Qi Liu, Hao Wang, Chengqiang Lu, Chee-Kong Lee
Predicting molecular properties with data-driven methods has drawn much attentio
n in recent years. Particularly, Graph Neural Networks (GNNs) have demonstrated
remarkable success in various molecular generation and prediction tasks. In case

s where labeled data is scarce, GNNs can be pre-trained on unlabeled molecular data to first learn the general semantic and structural information before being finetuned for specific tasks. However, most existing self-supervised pretraining frameworks for GNNs only focus on node-level or graph-level tasks. These approaches cannot capture the rich information in subgraphs or graph motifs. For example, functional groups (frequently-occurred subgraphs in molecular graphs) often carry indicative information about the molecular properties. To bridge this gap, we propose Motif-based Graph Self-supervised Learning (MGSSL) by introducing a novel self-supervised motif generation framework for GNNs. First, for motif extraction from molecular graphs, we design a molecule fragmentation method that leverages a retrosynthesis-based algorithm BRICS and additional rules for controlling the size of motif vocabulary. Second, we design a general motif-based generative pretraining framework in which GNNs are asked to make topological and label predictions. This generative framework can be implemented in two different ways, i.e., breadth-first or depth-first. Finally, to take the multi-scale information in molecular graphs into consideration, we introduce a multi-level self-supervised pre-training. Extensive experiments on various downstream benchmark tasks show that our methods outperform all state-of-the-art baselines.
****
On Inductive Biases for Heterogeneous Treatment Effect Estimation
Alicia Curth, Mihaela van der Schaar
We investigate how to exploit structural similarities of an individual's potential outcomes (POs) under different treatments to obtain better estimates of conditional average treatment effects in finite samples. Especially when it is unknown whether a treatment has an effect at all, it is natural to hypothesize that the POs are similar -- yet, some existing strategies for treatment effect estimation employ regularization schemes that implicitly encourage heterogeneity even when it does not exist and fail to fully make use of shared structure. In this paper, we investigate and compare three end-to-end learning strategies to overcome this problem -- based on regularization, reparametrization and a flexible multi-task architecture -- each encoding inductive bias favoring shared behavior across POs. To build understanding of their relative strengths, we implement all strategies using neural networks and conduct a wide range of semi-synthetic experiments. We observe that all three approaches can lead to substantial improvements upon numerous baselines and gain insight into performance differences across various experimental settings.
****
DP-SSL: Towards Robust Semi-supervised Learning with A Few Labeled Samples
Yi Xu, Jiandong Ding, Lu Zhang, Shuigeng Zhou
The scarcity of labeled data is a critical obstacle to deep learning. Semi-supervised learning (SSL) provides a promising way to leverage unlabeled data by pseudo labels. However, when the size of labeled data is very small (say a few labeled samples per class), SSL performs poorly and unstably, possibly due to the low quality of learned pseudo labels. In this paper, we propose a new SSL method called DP-SSL that adopts an innovative data programming (DP) scheme to generate probabilistic labels for unlabeled data. Different from existing DP methods that rely on human experts to provide initial labeling functions (LFs), we develop a multiple-choice learning~(MCL) based approach to automatically generate LFs from scratch in SSL style. With the noisy labels produced by the LFs, we design a label model to resolve the conflict and overlap among the noisy labels, and finally infer probabilistic labels for unlabeled samples. Extensive experiments on four standard SSL benchmarks show that DP-SSL can provide reliable labels for unlabeled data and achieve better classification performance on test sets than existing SSL methods, especially when only a small number of labeled samples are available. Concretely, for CIFAR-10 with only 40 labeled samples, DP-SSL achieves 93.82% annotation accuracy on unlabeled data and 93.46% classification accuracy on test data, which are higher than the SOTA results.
****
Transformer in Transformer
Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, 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 Graph Augmentation to Improve Graph Contrastive Learning
Susheel Suresh, Pan Li, Cong Hao, Jennifer Neville
Self-supervised learning of graph neural networks (GNN) is in great need because of the widespread label scarcity issue in real-world graph/network data. Graph contrastive learning (GCL), by training GNNs to maximize the correspondence between the representations of the same graph in its different augmented forms, may yield robust and transferable GNNs even without using labels. However, GNNs trained by traditional GCL often risk capturing redundant graph features and thus may be brittle and provide sub-par performance in downstream tasks. Here, we propose a novel principle, termed adversarial-GCL (\textit{AD-GCL}), which enables GNNs to avoid capturing redundant information during the training by optimizing adversarial graph augmentation strategies used in GCL. We pair AD-GCL with theoretical explanations and design a practical instantiation based on trainable edge-dropping graph augmentation. We experimentally validate AD-GCL by comparing with the state-of-the-art GCL methods and achieve performance gains of up-to~14\% in unsupervised, ~6\% in transfer and~3\% in semi-supervised learning settings overall with 18 different benchmark datasets for the tasks of molecule property regression and classification, and social network classification.
****
Online Control of Unknown Time-Varying Dynamical Systems
Edgar Minasyan, Paula Gradu, Max Simchowitz, Elad Hazan
We study online control of time-varying linear systems with unknown dynamics in the nonstochastic control model. At a high level, we demonstrate that this setting is \emph{qualitatively harder} than that of either unknown time-invariant or known time-varying dynamics, and complement our negative results with algorithmic upper bounds in regimes where sublinear regret is possible. More specifically, we study regret bounds with respect to common classes of policies: Disturbance Action (SLS), Disturbance Response (Youla), and linear feedback policies. While these three classes are essentially equivalent for LTI systems, we demonstrate that these equivalences break down for time-varying systems. We prove a lower bound that no algorithm can obtain sublinear regret with respect to the first two classes unless a certain measure of system variability also scales sublinearly in the horizon. Furthermore, we show that offline planning over the state linear feedback policies is NP-hard, suggesting hardness of the online learning problem. On the positive side, we give an efficient algorithm that attains a sublinear regret bound against the class of Disturbance Response policies up to the aforementioned system variability term. In fact, our algorithm enjoys sublinear \emph{adaptive} regret bounds, which is a strictly stronger metric than standard regret and is more appropriate for time-varying systems. We sketch extensions to Disturbance Action policies and partial observation, and propose an inefficient algorithm for regret against linear state feedback policies.
****
Contrastive Reinforcement Learning of Symbolic Reasoning Domains
Gabriel Poesia, WenXin Dong, Noah Goodman
Abstract symbolic reasoning, as required in domains such as mathematics and logic, is a key component of human intelligence. Solvers for these domains have important applications, especially to computer-assisted education. But learning to solve symbolic problems is challenging for machine learning algorithms. Existing models either learn from human solutions or use hand-engineered features, making them expensive to apply in new domains. In this paper, we instead consider symbolic domains as simple environments where states and actions are given as unstructured text, and binary rewards indicate whether a problem is solved. This flexible setup makes it easy to specify new domains, but search and planning become challenging. We introduce five environments inspired by the Mathematics Common Core Curriculum, and observe that existing Reinforcement Learning baselines perfor

m poorly. We then present a novel learning algorithm, Contrastive Policy Learnin g (ConPoLe) that explicitly optimizes the InfoNCE loss, which lower bounds the m utual information between the current state and next states that continue on a p ath to the solution. ConPoLe successfully solves all four domains. Moreover, pro blem representations learned by ConPoLe enable accurate prediction of the catego ries of problems in a real mathematics curriculum. Our results suggest new direc tions for reinforcement learning in symbolic domains, as well as applications to mathematics education.
****
Spatial Ensemble: a Novel Model Smoothing Mechanism for Student-Teacher Framewor k
Tengteng Huang, Yifan Sun, Xun Wang, Haotian Yao, Chi Zhang
Model smoothing is of central importance for obtaining a reliable teacher model in the student-teacher framework, where the teacher generates surrogate supervis ion signals to train the student. A popular model smoothing method is the Tempor al Moving Average (TMA), which continuously averages the teacher parameters with the up-to-date student parameters. In this paper, we propose ''Spatial Ensemble '', a novel model smoothing mechanism in parallel with TMA. Spatial Ensemble ran domly picks up a small fragment of the student model to directly replace the cor responding fragment of the teacher model. Consequentially, it stitches different fragments of historical student models into a unity, yielding the ''Spatial Ens emble'' effect. Spatial Ensemble obtains comparable student-teacher learning per formance by itself and demonstrates valuable complementarity with temporal movin g average. Their integration, named Spatial-Temporal Smoothing, brings general (sometimes significant) improvement to the student-teacher learning framework on a variety of state-of-the-art methods. For example, based on the self-supervised method BYOL, it yields +0.9% top-1 accuracy improvement on ImageNet, while base d on the semi-supervised approach FixMatch, it increases the top-1 accuracy by a round +6% on CIFAR-10 when only few training labels are available. Codes and mod els are available at: https://github.com/tengteng95/Spatial_Ensemble.
****
Probabilistic Tensor Decomposition of Neural Population Spiking Activity
Hugo Soulat, Sepiedeh Keshavarzi, Troy Margrie, Maneesh Sahani
The firing of neural populations is coordinated across cells, in time, and acros s experimentalconditions or repeated experimental trials; and so a full understa nding of the computationalsignificance of neural responses must be based on a se paration of these different contributions tostructured activity.Tensor decomposi tion is an approach to untangling the influence of multiple factors in data that iscommon in many fields. However, despite some recent interest in neuroscience , wider applicabilityof the approach is hampered by the lack of a full probabili stic treatment allowing principledinference of a decomposition from non-Gaussian spike-count data.Here, we extend the Pólya-Gamma (PG) augmentation, previously used in sampling-based Bayesianinference, to implement scalable variational infe rence in non-conjugate spike-count models.Using this new approach, we develop te chniques related to automatic relevance determination to inferthe most appropria te tensor rank, as well as to incorporate priors based on known brain anatomy su chas the segregation of cell response properties by brain area.We apply the mode l to neural recordings taken under conditions of visual-vestibular sensoryintegr ation, revealing how the encoding of self- and visual-motion signals is modulate d by thesensory information available to the animal.
****
Recurrent Bayesian Classifier Chains for Exact Multi-Label Classification
Walter Gerych, Tom Hartvigsen, Luke Buquicchio, Emmanuel Agu, Elke A. Rundenstei ner
Exact multi-label classification is the task of assigning each datapoint a set o f class labels such that the assigned set exactly matches the ground truth. Opti mizing for exact multi-label classification is important in domains where missin g a single label can be especially costly, such as in object detection for auton omous vehicles or symptom classification for disease diagnosis. Recurrent Classi fier Chains (RCCs), a recurrent neural network extension of ensemble-based class

ifier chains, are the state-of-the-art exact multi-label classification method for maximizing subset accuracy. However, RCCs iteratively predict classes with an unprincipled ordering, and therefore indiscriminately condition class probabilities. These disadvantages make RCCs prone to predicting inaccurate label sets. In this work we propose Recurrent Bayesian Classifier Chains (RBCCs), which learn a Bayesian network of class dependencies and leverage this network in order to condition the prediction of child nodes only on their parents. By conditioning predictions in this way, we perform principled and non-noisy class prediction. We demonstrate the effectiveness of our RBCC method on a variety of real-world multi-label datasets, where we routinely outperform the state of the art methods for exact multi-label classification.
****
Wasserstein Flow Meets Replicator Dynamics: A Mean-Field Analysis of Representation Learning in Actor-Critic
Yufeng Zhang, Siyu Chen, Zhuoran Yang, Michael Jordan, Zhaoran Wang
Actor-critic (AC) algorithms, empowered by neural networks, have had significant empirical success in recent years. However, most of the existing theoretical support for AC algorithms focuses on the case of linear function approximations, or linearized neural networks, where the feature representation is fixed throughout training. Such a limitation fails to capture the key aspect of representation learning in neural AC, which is pivotal in practical problems. In this work, we take a mean-field perspective on the evolution and convergence of feature-based neural AC. Specifically, we consider a version of AC where the actor and critic are represented by overparameterized two-layer neural networks and are updated with two-timescale learning rates. The critic is updated by temporal-difference (TD) learning with a larger stepsize while the actor is updated via proximal policy optimization (PPO) with a smaller stepsize. In the continuous-time and infinite-width limiting regime, when the timescales are properly separated, we prove that neural AC finds the globally optimal policy at a sublinear rate. Additionally, we prove that the feature representation induced by the critic network is allowed to evolve within a neighborhood of the initial one.
****
Assessing Fairness in the Presence of Missing Data
Yiliang Zhang, Qi Long
Missing data are prevalent and present daunting challenges in real data analysis. While there is a growing body of literature on fairness in analysis of fully observed data, there has been little theoretical work on investigating fairness in analysis of incomplete data. In practice, a popular analytical approach for dealing with missing data is to use only the set of complete cases, i.e., observations with all features fully observed to train a prediction algorithm. However, depending on the missing data mechanism, the distribution of complete cases and the distribution of the complete data may be substantially different. When the goal is to develop a fair algorithm in the complete data domain where there are no missing values, an algorithm that is fair in the complete case domain may show disproportionate bias towards some marginalized groups in the complete data domain. To fill this significant gap, we study the problem of estimating fairness in the complete data domain for an arbitrary model evaluated merely using complete cases. We provide upper and lower bounds on the fairness estimation error and conduct numerical experiments to assess our theoretical results. Our work provides the first known theoretical results on fairness guarantee in analysis of incomplete data.
****
Adversarial Attack Generation Empowered by Min-Max Optimization
Jingkang Wang, Tianyun Zhang, Sijia Liu, Pin-Yu Chen, Jiacen Xu, Makan Fardad, Bo Li
The worst-case training principle that minimizes the maximal adversarial loss, also known as adversarial training (AT), has shown to be a state-of-the-art approach for enhancing adversarial robustness. Nevertheless, min-max optimization beyond the purpose of AT has not been rigorously explored in the adversarial context. In this paper, we show how a general notion of min-max optimization over mult

iple domains can be leveraged to the design of different types of adversarial attacks. In particular, given a set of risk sources, minimizing the worst-case attack loss can be reformulated as a min-max problem by introducing domain weights that are maximized over the probability simplex of the domain set. We showcase this unified framework in three attack generation problems -- attacking model ensembles, devising universal perturbation under multiple inputs, and crafting attacks resilient to data transformations. Extensive experiments demonstrate that our approach leads to substantial attack improvement over the existing heuristic strategies as well as robustness improvement over state-of-the-art defense methods against multiple perturbation types. Furthermore, we find that the self-adjusted domain weights learned from min-max optimization can provide a holistic tool to explain the difficulty level of attack across domains.
****
Safe Pontryagin Differentiable Programming
Wanxin Jin, Shaoshuai Mou, George J. Pappas
We propose a Safe Pontryagin Differentiable Programming (Safe PDP) methodology, which establishes a theoretical and algorithmic framework to solve a broad class of safety-critical learning and control tasks---problems that require the guarantee of safety constraint satisfaction at any stage of the learning and control progress. In the spirit of interior-point methods, Safe PDP handles different types of system constraints on states and inputs by incorporating them into the cost or loss through barrier functions. We prove three fundamentals of the proposed Safe PDP: first, both the solution and its gradient in the backward pass can be approximated by solving their more efficient unconstrained counterparts; second, the approximation for both the solution and its gradient can be controlled for arbitrary accuracy by a barrier parameter; and third, importantly, all intermediate results throughout the approximation and optimization strictly respect the constraints, thus guaranteeing safety throughout the entire learning and control process. We demonstrate the capabilities of Safe PDP in solving various safety-critical tasks, including safe policy optimization, safe motion planning, and learning MPCs from demonstrations, on different challenging systems such as 6-DoF maneuvering quadrotor and 6-DoF rocket powered landing.
****
Class-Disentanglement and Applications in Adversarial Detection and Defense
Kaiwen Yang, Tianyi Zhou, Yonggang Zhang, Xinmei Tian, Dacheng Tao
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Active 3D Shape Reconstruction from Vision and Touch
Edward Smith, David Meger, Luis Pineda, Roberto Calandra, Jitendra Malik, Adriana Romero Soriano, Michal Drozdzal
Humans build 3D understandings of the world through active object exploration, using jointly their senses of vision and touch. However, in 3D shape reconstruction, most recent progress has relied on static datasets of limited sensory data such as RGB images, depth maps or haptic readings, leaving the active exploration of the shape largely unexplored. In active touch sensing for 3D reconstruction, the goal is to actively select the tactile readings that maximize the improvement in shape reconstruction accuracy. However, the development of deep learning-based active touch models is largely limited by the lack of frameworks for shape exploration. In this paper, we focus on this problem and introduce a system composed of: 1) a haptic simulator leveraging high spatial resolution vision-based tactile sensors for active touching of 3D objects; 2) a mesh-based 3D shape reconstruction model that relies on tactile or visuotactile signals; and 3) a set of data-driven solutions with either tactile or visuotactile priors to guide the shape exploration. Our framework enables the development of the first fully data-driven solutions to active touch on top of learned models for object understanding. Our experiments show the benefits of such solutions in the task of 3D shape u

nderstanding where our models consistently outperform natural baselines. We prov
ide our framework as a tool to foster future research in this direction.
****
CAPE: Encoding Relative Positions with Continuous Augmented Positional Embedding
s
Tatiana Likhomanenko, Qiantong Xu, Gabriel Synnaeve, Ronan Collobert, Alex Rogoz
hnikov
Without positional information, attention-based Transformer neural networks are
permutation-invariant. Absolute or relative positional embeddings are the most p
opular ways to feed Transformer models with positional information. Absolute pos
itional embeddings are simple to implement, but suffer from generalization issue
s when evaluating on sequences longer than seen at training time. Relative posit
ions are more robust to input length change, but are more complex to implement a
nd yield inferior model throughput due to extra computational and memory costs.
In this paper, we propose an augmentation-based approach (CAPE) for absolute pos
itional embeddings, which keeps the advantages of both absolute (simplicity and
speed) and relative positional embeddings (better generalization). In addition,
our empirical evaluation on state-of-the-art models in machine translation, imag
e and speech recognition demonstrates that CAPE leads to better generalization p
erformance as well as increased stability with respect to training hyper-paramet
ers.
****
Multi-armed Bandit Requiring Monotone Arm Sequences
Ningyuan 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.
****
Gradient Driven Rewards to Guarantee Fairness in Collaborative Machine Learning
Xinyi Xu, Lingjuan Lyu, Xingjun Ma, Chenglin Miao, Chuan Sheng Foo, Bryan Kian H
siang Low
In collaborative machine learning(CML), multiple agents pool their resources(e.g
., data) together for a common learning task. In realistic CML settings where th
e agents are self-interested and not altruistic, they may be unwilling to share
data or model information without adequate rewards. Furthermore, as the data/mod
el information shared by the agents may differ in quality, designing rewards whi
ch are fair to them is important so that they would not feel exploited nor disco
uraged from sharing. In this paper, we adopt federated learning as the CML parad
igm, propose a novel cosine gradient Shapley value(CGSV) to fairly evaluate the
expected marginal contribution of each agent’s uploaded model parameter update/g
radient without needing an auxiliary validation dataset, and based on the CGSV,
design a novel training-time gradient reward mechanism with a fairness guarantee
 by sparsifying the aggregated parameter update/gradient downloaded from the ser
ver as reward to each agent such that its resulting quality is commensurate to t
hat of the agent’s uploaded parameter update/gradient. We empirically demonstrat
e the effectiveness of our fair gradient reward mechanism on multiple benchmark
datasets in terms of fairness, predictive performance, and time overhead.
****
Generalizable Imitation Learning from Observation via Inferring Goal Proximity
Youngwoon Lee, Andrew Szot, Shao-Hua Sun, Joseph J. Lim
Task progress is intuitive and readily available task information that can guide
 an agent closer to the desired goal. Furthermore, a task progress estimator can
 generalize to new situations. From this intuition, we propose a simple yet effe
ctive imitation learning from observation method for a goal-directed task using
a learned goal proximity function as a task progress estimator for better genera
lization to unseen states and goals. We obtain this goal proximity function from
 expert demonstrations and online agent experience, and then use the learned goa
l proximity as a dense reward for policy training. We demonstrate that our propo
sed method can robustly generalize compared to prior imitation learning methods

on a set of goal-directed tasks in navigation, locomotion, and robotic manipulation, even with demonstrations that cover only a part of the states.
****
DualNet: Continual Learning, Fast and Slow
Quang Pham, Chenghao Liu, Steven Hoi
According to Complementary Learning Systems (CLS) theory~\cite{mcclelland1995there} in neuroscience, humans do effective \emph{continual learning} through two complementary systems: a fast learning system centered on the hippocampus for rapid learning of the specifics and individual experiences, and a slow learning system located in the neocortex for the gradual acquisition of structured knowledge about the environment. Motivated by this theory, we propose a novel continual learning framework named ``DualNet", which comprises a fast learning system for supervised learning of pattern-separated representation from specific tasks and a slow learning system for unsupervised representation learning of task-agnostic general representation via a Self-Supervised Learning (SSL) technique. The two fast and slow learning systems are complementary and work seamlessly in a holistic continual learning framework. Our extensive experiments on two challenging continual learning benchmarks of CORE50 and miniImageNet show that DualNet outperforms state-of-the-art continual learning methods by a large margin. We further conduct ablation studies of different SSL objectives to validate DualNet's efficacy, robustness, and scalability. Code is publicly available at \url{https://github.com/phquang/DualNet}.
****
Deformable Butterfly: A Highly Structured and Sparse Linear Transform
Rui Lin, Jie Ran, King Hung Chiu, Graziano Chesi, Ngai Wong
We introduce a new kind of linear transform named Deformable Butterfly (DeBut) that generalizes the conventional butterfly matrices and can be adapted to various input-output dimensions. It inherits the fine-to-coarse-grained learnable hierarchy of traditional butterflies and when deployed to neural networks, the prominent structures and sparsity in a DeBut layer constitutes a new way for network compression. We apply DeBut as a drop-in replacement of standard fully connected and convolutional layers, and demonstrate its superiority in homogenizing a neural network and rendering it favorable properties such as light weight and low inference complexity, without compromising accuracy. The natural complexity-accuracy tradeoff arising from the myriad deformations of a DeBut layer also opens up new rooms for analytical and practical research. The codes and Appendix are publicly available at: https://github.com/ruilin0212/DeBut.
****
Why Do Pretrained Language Models Help in Downstream Tasks? An Analysis of Head and Prompt Tuning
Colin Wei, Sang Michael Xie, Tengyu Ma
Pretrained language models have achieved state-of-the-art performance when adapted to a downstream NLP task. However, theoretical analysis of these models is scarce and challenging since the pretraining and downstream tasks can be very different. We propose an analysis framework that links the pretraining and downstream tasks with an underlying latent variable generative model of text -- the downstream classifier must recover a function of the posterior distribution over the latent variables. We analyze head tuning (learning a classifier on top of the frozen pretrained model) and prompt tuning in this setting. The generative model in our analysis is either a Hidden Markov Model (HMM) or an HMM augmented with a latent memory component, motivated by long-term dependencies in natural language. We show that 1) under certain non-degeneracy conditions on the HMM, simple classification heads can solve the downstream task, 2) prompt tuning obtains downstream guarantees with weaker non-degeneracy conditions, and 3) our recovery guarantees for the memory-augmented HMM are stronger than for the vanilla HMM because task-relevant information is easier to recover from the long-term memory. Experiments on synthetically generated data from HMMs back our theoretical findings.
****
Learning Diverse Policies in MOBA Games via Macro-Goals
Yiming Gao, Bei Shi, Xueying Du, Liang Wang, Guangwei Chen, Zhenjie Lian, Fuhao

Qiu, GUOAN HAN, Weixuan Wang, Deheng Ye, Qiang Fu, Wei Yang, Lanxiao Huang
Recently, many researchers have made successful progress in building the AI systems for MOBA-game-playing with deep reinforcement learning, such as on Dota 2 and Honor of Kings. Even though these AI systems have achieved or even exceeded human-level performance, they still suffer from the lack of policy diversity. In this paper, we propose a novel Macro-Goals Guided framework, called MGG, to learn diverse policies in MOBA games. MGG abstracts strategies as macro-goals from human demonstrations and trains a Meta-Controller to predict these macro-goals. To enhance policy diversity, MGG samples macro-goals from the Meta-Controller prediction and guides the training process towards these goals. Experimental results on the typical MOBA game Honor of Kings demonstrate that MGG can execute diverse policies in different matches and lineups, and also outperform the state-of-the-art methods over 102 heroes.
****
Evaluation of Human-AI Teams for Learned and Rule-Based Agents in Hanabi
Ho Chit Siu, Jaime Peña, Edenna Chen, Yutai Zhou, Victor Lopez, Kyle Palko, Kimberlee Chang, Ross Allen
Deep reinforcement learning has generated superhuman AI in competitive games such as Go and StarCraft. Can similar learning techniques create a superior AI teammate for human-machine collaborative games? Will humans prefer AI teammates that improve objective team performance or those that improve subjective metrics of trust? In this study, we perform a single-blind evaluation of teams of humans and AI agents in the cooperative card game Hanabi, with both rule-based and learning-based agents. In addition to the game score, used as an objective metric of the human-AI team performance, we also quantify subjective measures of the human's perceived performance, teamwork, interpretability, trust, and overall preference of AI teammate. We find that humans have a clear preference toward a rule-based AI teammate (SmartBot) over a state-of-the-art learning-based AI teammate (Other-Play) across nearly all subjective metrics, and generally view the learning-based agent negatively, despite no statistical difference in the game score. This result has implications for future AI design and reinforcement learning benchmarking, highlighting the need to incorporate subjective metrics of human-AI teaming rather than a singular focus on objective task performance.
****
Counterfactual Invariance to Spurious Correlations in Text Classification
Victor Veitch, Alexander D'Amour, Steve Yadlowsky, Jacob Eisenstein
Informally, a 'spurious correlation' is the dependence of a model on some aspect of the input data that an analyst thinks shouldn't matter. In machine learning, these have a know-it-when-you-see-it character; e.g., changing the gender of a sentence's subject changes a sentiment predictor's output. To check for spurious correlations, we can 'stress test' models by perturbing irrelevant parts of input data and seeing if model predictions change. In this paper, we study stress testing using the tools of causal inference. We introduce counterfactual invariance as a formalization of the requirement that changing irrelevant parts of the input shouldn't change model predictions. We connect counterfactual invariance to out-of-domain model performance, and provide practical schemes for learning (approximately) counterfactual invariant predictors (without access to counterfactual examples). It turns out that both the means and implications of counterfactual invariance depend fundamentally on the true underlying causal structure of the data---in particular, whether the label causes the features or the features cause the label. Distinct causal structures require distinct regularization schemes to induce counterfactual invariance. Similarly, counterfactual invariance implies different domain shift guarantees depending on the underlying causal structure. This theory is supported by empirical results on text classification.
****
Better Safe Than Sorry: Preventing Delusive Adversaries with Adversarial Training
Lue Tao, Lei Feng, Jinfeng Yi, Sheng-Jun Huang, Songcan 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.
****
Determinantal point processes based on orthogonal polynomials for sampling minibatches in SGD
Rémi Bardenet, Subhroshekhar Ghosh, Meixia 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-authors prior to requesting a name change in the electronic proceedings.
****
Revisiting Contrastive Methods for Unsupervised Learning of Visual Representations
Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Luc V Gool
Contrastive self-supervised learning has outperformed supervised pretraining on many downstream tasks like segmentation and object detection. However, current methods are still primarily applied to curated datasets like ImageNet. In this paper, we first study how biases in the dataset affect existing methods. Our results show that an approach like MoCo works surprisingly well across: (i) object- versus scene-centric, (ii) uniform versus long-tailed and (iii) general versus domain-specific datasets. Second, given the generality of the approach, we try to realize further gains with minor modifications. We show that learning additional invariances - through the use of multi-scale cropping, stronger augmentations and nearest neighbors - improves the representations. Finally, we observe that MoCo learns spatially structured representations when trained with a multi-crop strategy. The representations can be used for semantic segment retrieval and video instance segmentation without finetuning. Moreover, the results are on par with specialized models. We hope this work will serve as a useful study for other researchers.
****
Neural Analysis and Synthesis: Reconstructing Speech from Self-Supervised Representations
Hyeong-Seok Choi, Juheon Lee, Wansoo Kim, Jie Lee, Hoon Heo, Kyogu Lee
We present a neural analysis and synthesis (NANSY) framework that can manipulate the voice, pitch, and speed of an arbitrary speech signal. Most of the previous works have focused on using information bottleneck to disentangle analysis features for controllable synthesis, which usually results in poor reconstruction quality. We address this issue by proposing a novel training strategy based on information perturbation. The idea is to perturb information in the original input signal (e.g., formant, pitch, and frequency response), thereby letting synthesis networks selectively take essential attributes to reconstruct the input signal. Because NANSY does not need any bottleneck structures, it enjoys both high reconstruction quality and controllability. Furthermore, NANSY does not require any labels associated with speech data such as text and speaker information, but rather uses a new set of analysis features, i.e., wav2vec feature and newly proposed pitch feature, Yingram, which allows for fully self-supervised training. Taking advantage of fully self-supervised training, NANSY can be easily extended to a multilingual setting by simply training it with a multilingual dataset. The experiments show that NANSY can achieve significant improvement in performance in several applications such as zero-shot voice conversion, pitch shift, and time-scale modification.
****
Auto-Encoding Knowledge Graph for Unsupervised Medical Report Generation
Fenglin Liu, Chenyu You, Xian Wu, Shen Ge, Sheng wang, Xu Sun
Medical report generation, which aims to automatically generate a long and coherent report of a given medical image, has been receiving growing research interests. Existing approaches mainly adopt a supervised manner and heavily rely on coupled image-report pairs. However, in the medical domain, building a large-scale image-report paired dataset is both time-consuming and expensive. To relax the dependency on paired data, we propose an unsupervised model Knowledge Graph Auto-

Encoder (KGAE) which accepts independent sets of images and reports in training. KGAE consists of a pre-constructed knowledge graph, a knowledge-driven encoder and a knowledge-driven decoder. The knowledge graph works as the shared latent space to bridge the visual and textual domains; The knowledge-driven encoder projects medical images and reports to the corresponding coordinates in this latent space and the knowledge-driven decoder generates a medical report given a coordinate in this space. Since the knowledge-driven encoder and decoder can be trained with independent sets of images and reports, KGAE is unsupervised. The experiments show that the unsupervised KGAE generates desirable medical reports without using any image-report training pairs. Moreover, KGAE can also work in both semi-supervised and supervised settings, and accept paired images and reports in training. By further fine-tuning with image-report pairs, KGAE consistently outperforms the current state-of-the-art models on two datasets.
****

Diffusion Normalizing Flow
Qinsheng Zhang, Yongxin Chen
We present a novel generative modeling method called diffusion normalizing flow based on stochastic differential equations (SDEs). The algorithm consists of two neural SDEs: a forward SDE that gradually adds noise to the data to transform the data into Gaussian random noise, and a backward SDE that gradually removes the noise to sample from the data distribution. By jointly training the two neural SDEs to minimize a common cost function that quantifies the difference between the two, the backward SDE converges to a diffusion process the starts with a Gaussian distribution and ends with the desired data distribution. Our method is closely related to normalizing flow and diffusion probabilistic models, and can be viewed as a combination of the two. Compared with normalizing flow, diffusion normalizing flow is able to learn distributions with sharp boundaries. Compared with diffusion probabilistic models, diffusion normalizing flow requires fewer discretization steps and thus has better sampling efficiency. Our algorithm demonstrates competitive performance in both high-dimension data density estimation and image generation tasks.
****

Introspective Distillation for Robust Question Answering
Yulei Niu, Hanwang Zhang
Question answering (QA) models are well-known to exploit data bias, e.g., the language prior in visual QA and the position bias in reading comprehension. Recent debiasing methods achieve good out-of-distribution (OOD) generalizability with a considerable sacrifice of the in-distribution (ID) performance. Therefore, they are only applicable in domains where the test distribution is known in advance. In this paper, we present a novel debiasing method called Introspective Distillation (IntroD) to make the best of both worlds for QA. Our key technical contribution is to blend the inductive bias of OOD and ID by introspecting whether a training sample fits in the factual ID world or the counterfactual OOD one. Experiments on visual QA datasets VQA v2, VQA-CP, and reading comprehension dataset SQuAD demonstrate that our proposed IntroD maintains the competitive OOD performance compared to other debiasing methods, while sacrificing little or even achieving better ID performance compared to the non-debiasing ones.
****

Rethinking the Pruning Criteria for Convolutional Neural Network
Zhongzhan Huang, Wenqi Shao, Xinjiang Wang, Liang Lin, Ping 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.
****

Adaptive Machine Unlearning
Varun Gupta, Christopher Jung, Seth Neel, Aaron Roth, Saeed Sharifi-Malvajerdi, Chris Waites
Data deletion algorithms aim to remove the influence of deleted data points from trained models at a cheaper computational cost than fully retraining those mode

ls. However, for sequences of deletions, most prior work in the non-convex setting gives valid guarantees only for sequences that are chosen independently of the models that are published. If people choose to delete their data as a function of the published models (because they don’t like what the models reveal about them, for example), then the update sequence is adaptive. In this paper, we give a general reduction from deletion guarantees against adaptive sequences to deletion guarantees against non-adaptive sequences, using differential privacy and its connection to max information. Combined with ideas from prior work which give guarantees for non-adaptive deletion sequences, this leads to extremely flexible algorithms able to handle arbitrary model classes and training methodologies, giving strong provable deletion guarantees for adaptive deletion sequences. We show in theory how prior work for non-convex models fails against adaptive deletion sequences, and use this intuition to design a practical attack against the SISA algorithm of Bourtoule et al. [2021] on CIFAR-10, MNIST, Fashion-MNIST.
****
EditGAN: High-Precision Semantic Image Editing
Huan Ling, Karsten Kreis, Daiqing Li, Seung Wook Kim, Antonio Torralba, Sanja Fidler
Generative adversarial networks (GANs) have recently found applications in image editing. However, most GAN-based image editing methods often require large-scale datasets with semantic segmentation annotations for training, only provide high-level control, or merely interpolate between different images. Here, we propose EditGAN, a novel method for high-quality, high-precision semantic image editing, allowing users to edit images by modifying their highly detailed part segmentation masks, e.g., drawing a new mask for the headlight of a car. EditGAN builds on a GAN framework that jointly models images and their semantic segmentation, requiring only a handful of labeled examples – making it a scalable tool for editing. Specifically, we embed an image into the GAN’s latent space and perform conditional latent code optimization according to the segmentation edit, which effectively also modifies the image. To amortize optimization, we find “editing vectors” in latent space that realize the edits. The framework allows us to learn an arbitrary number of editing vectors, which can then be directly applied on other images at interactive rates. We experimentally show that EditGAN can manipulate images with an unprecedented level of detail and freedom while preserving full image quality. We can also easily combine multiple edits and perform plausible edits beyond EditGAN’s training data. We demonstrate EditGAN on a wide variety of image types and quantitatively outperform several previous editing methods on standard editing benchmark tasks.
****
Deep Molecular Representation Learning via Fusing Physical and Chemical Information
Shuwen Yang, Ziyao Li, Guojie Song, Lingsheng Cai
Molecular representation learning is the first yet vital step in combining deep learning and molecular science. To push the boundaries of molecular representation learning, we present PhysChem, a novel neural architecture that learns molecular representations via fusing physical and chemical information of molecules. PhysChem is composed of a physicist network (PhysNet) and a chemist network (ChemNet). PhysNet is a neural physical engine that learns molecular conformations through simulating molecular dynamics with parameterized forces; ChemNet implements geometry-aware deep message-passing to learn chemical / biomedical properties of molecules. Two networks specialize in their own tasks and cooperate by providing expertise to each other. By fusing physical and chemical information, PhysChem achieved state-of-the-art performances on MoleculeNet, a standard molecular machine learning benchmark. The effectiveness of PhysChem was further corroborated on cutting-edge datasets of SARS-CoV-2.
****
Neural optimal feedback control with local learning rules
Johannes Friedrich, Siavash Golkar, Shiva Farashahi, Alexander Genkin, Anirvan Sengupta, Dmitri Chklovskii
A major problem in motor control is understanding how the brain plans and execut

es proper movements in the face of delayed and noisy stimuli. A prominent framew
ork for addressing such control problems is Optimal Feedback Control (OFC). OFC
generates control actions that optimize behaviorally relevant criteria by integr
ating noisy sensory stimuli and the predictions of an internal model using the K
alman filter or its extensions. However, a satisfactory neural model of Kalman f
iltering and control is lacking because existing proposals have the following l
imitations: not considering the delay of sensory feedback, training in alternati
ng phases, requiring knowledge of the noise covariance matrices, as well as that
 of systems dynamics. Moreover, the majority of these studies considered Kalman
filtering in isolation, and not jointly with control. To address these shortcomi
ngs, we introduce a novel online algorithm which combines adaptive Kalman filter
ing with a model free control approach (i.e., policy gradient algorithm). We im
plement this algorithm in a biologically plausible neural network with local syn
aptic plasticity rules. This network, with local synaptic plasticity rules, perf
orms system identification, Kalman filtering and control with delayed noisy sens
ory feedback. This network performs system identification and Kalman filtering,
without the need for multiple phases with distinct update rules or the knowledge
 of the noise covariances. It can perform state estimation with delayed sensory
 feedback, with the help of an internal model. It learns the control policy with
out requiring any knowledge of the dynamics, thus avoiding the need for weight t
ransport. In this way, our implementation of OFC solves the credit assignment pr
oblem needed to produce the appropriate sensory-motor control in the presence of
 stimulus delay.
****
Reinforcement Learning in Linear MDPs: Constant Regret and Representation Select
ion
Matteo Papini, Andrea Tirinzoni, Aldo Pacchiano, Marcello Restelli, Alessandro L
azaric, Matteo Pirotta
We study the role of the representation of state-action value functions in regre
t minimization in finite-horizon Markov Decision Processes (MDPs) with linear st
ructure. We first derive a necessary condition on the representation, called uni
versally spanning optimal features (UNISOFT), to achieve constant regret in any
MDP with linear reward function. This result encompasses the well-known settings
 of low-rank MDPs and, more generally, zero inherent Bellman error (also known a
s the Bellman closure assumption). We then demonstrate that this condition is al
so sufficient for these classes of problems by deriving a constant regret bound
for two optimistic algorithms (LSVI-UCB and ELEANOR). Finally, we propose an alg
orithm for representation selection and we prove that it achieves constant regre
t when one of the given representations, or a suitable combination of them, sati
sfies the UNISOFT condition.
****
Noether Networks: meta-learning useful conserved quantities
Ferran Alet, Dylan Doblar, Allan Zhou, Josh Tenenbaum, Kenji Kawaguchi, Chelsea
Finn
Progress in machine learning (ML) stems from a combination of data availability,
 computational resources, and an appropriate encoding of inductive biases. Usefu
l biases often exploit symmetries in the prediction problem, such as convolution
al networks relying on translation equivariance. Automatically discovering these
 useful symmetries holds the potential to greatly improve the performance of ML
systems, but still remains a challenge. In this work, we focus on sequential pre
diction problems and take inspiration from Noether's theorem to reduce the probl
em of finding inductive biases to meta-learning useful conserved quantities. We
propose Noether Networks: a new type of architecture where a meta-learned conser
vation loss is optimized inside the prediction function. We show, theoretically
and experimentally, that Noether Networks improve prediction quality, providing
a general framework for discovering inductive biases in sequential problems.
****
Uncertainty-Driven Loss for Single Image Super-Resolution
Qian Ning, Weisheng Dong, Xin Li, Jinjian Wu, GUANGMING Shi
In low-level vision such as single image super-resolution (SISR), traditional MS

E or L1 loss function treats every pixel equally with the assumption that the im
portance of all pixels is the same. However, it has been long recognized that te
xture and edge areas carry more important visual information than smooth areas i
n photographic images. How to achieve such spatial adaptation in a principled ma
nner has been an open problem in both traditional model-based and modern learnin
g-based approaches toward SISR. In this paper, we propose a new adaptive weighte
d loss for SISR to train deep networks focusing on challenging situations such a
s textured and edge pixels with high uncertainty. Specifically, we introduce var
iance estimation characterizing the uncertainty on a pixel-by-pixel basis into S
ISR solutions so the targeted pixels in a high-resolution image (mean) and their
 corresponding uncertainty (variance) can be learned simultaneously. Moreover, u
ncertainty estimation allows us to leverage conventional wisdom such as sparsity
 prior for regularizing SISR solutions. Ultimately, pixels with large certainty
(e.g., texture and edge pixels) will be prioritized for SISR according to their
importance to visual quality. For the first time, we demonstrate that such uncer
tainty-driven loss can achieve better results than MSE or L1 loss for a wide ran
ge of network architectures. Experimental results on three popular SISR networks
 show that our proposed uncertainty-driven loss has achieved better PSNR perform
ance than traditional loss functions without any increased computation during te
sting. The code is available at https://see.xidian.edu.cn/faculty/wsdong/Project
s/UDL-SR.htm
****
GradInit: Learning to Initialize Neural Networks for Stable and Efficient Traini
ng
Chen Zhu, Renkun Ni, Zheng Xu, Kezhi Kong, W. Ronny Huang, Tom Goldstein
Innovations in neural architectures have fostered significant breakthroughs in l
anguage modeling and computer vision. Unfortunately, novel architectures often r
esult in challenging hyper-parameter choices and training instability if the net
work parameters are not properly initialized. A number of architecture-specific
initialization schemes have been proposed, but these schemes are not always port
able to new architectures. This paper presents GradInit, an automated and archit
ecture agnostic method for initializing neural networks. GradInit is based on a
simple heuristic; the norm of each network layer is adjusted so that a single st
ep of SGD or Adam with prescribed hyperparameters results in the smallest possib
le loss value. This adjustment is done by introducing a scalar multiplier variab
le in front of each parameter block, and then optimizing these variables using a
 simple numerical scheme. GradInit accelerates the convergence and test performa
nce of many convolutional architectures, both with or without skip connections,
and even without normalization layers. It also improves the stability of the ori
ginal Transformer architecture for machine translation, enabling training it wit
hout learning rate warmup using either Adam or SGD under a wide range of learnin
g rates and momentum coefficients. Code is available at https://github.com/zhuch
en03/gradinit.
****
Capacity and Bias of Learned Geometric Embeddings for Directed Graphs
Michael Boratko, Dongxu Zhang, Nicholas Monath, Luke Vilnis, Kenneth L Clarkson,
 Andrew McCallum
A wide variety of machine learning tasks such as knowledge base completion, onto
logy alignment, and multi-label classification can benefit from incorporating in
to learning differentiable representations of graphs or taxonomies. While vecto
rs in Euclidean space can theoretically represent any graph, much recent work sh
ows that alternatives such as complex, hyperbolic, order, or box embeddings have
 geometric properties better suited to modeling real-world graphs. Experimentall
y these gains are seen only in lower dimensions, however, with performance benef
its diminishing in higher dimensions. In this work, we introduce a novel variant
 of box embeddings that uses a learned smoothing parameter to achieve better rep
resentational capacity than vector models in low dimensions, while also avoiding
 performance saturation common to other geometric models in high dimensions. Fur
ther, we present theoretical results that prove box embeddings can represent any
 DAG. We perform rigorous empirical evaluations of vector, hyperbolic, and regio

n-based geometric representations on several families of synthetic and real-world directed graphs. Analysis of these results exposes correlations between different families of graphs, graph characteristics, model size, and embedding geometry, providing useful insights into the inductive biases of various differentiable graph representations.
****
Online Learning Of Neural Computations From Sparse Temporal Feedback
Lukas Braun, Tim Vogels
Neuronal computations depend on synaptic connectivity and intrinsic electrophysiological properties. Synaptic connectivity determines which inputs from presynaptic neurons are integrated, while cellular properties determine how inputs are filtered over time. Unlike their biological counterparts, most computational approaches to learning in simulated neural networks are limited to changes in synaptic connectivity. However, if intrinsic parameters change, neural computations are altered drastically. Here, we include the parameters that determine the intrinsic properties, e.g., time constants and reset potential, into the learning paradigm. Using sparse feedback signals that indicate target spike times, and gradient-based parameter updates, we show that the intrinsic parameters can be learned along with the synaptic weights to produce specific input-output functions. Specifically, we use a teacher-student paradigm in which a randomly initialised leaky integrate-and-fire or resonate-and-fire neuron must recover the parameters of a teacher neuron. We show that complex temporal functions can be learned online and without backpropagation through time, relying on event-based updates only. Our results are a step towards online learning of neural computations from ungraded and unsigned sparse feedback signals with a biologically inspired learning mechanism.
****
Self-Supervised Learning with Data Augmentations Provably Isolates Content from Style
Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, Francesco Locatello
Self-supervised representation learning has shown remarkable success in a number of domains. A common practice is to perform data augmentation via hand-crafted transformations intended to leave the semantics of the data invariant. We seek to understand the empirical success of this approach from a theoretical perspective. We formulate the augmentation process as a latent variable model by postulating a partition of the latent representation into a content component, which is assumed invariant to augmentation, and a style component, which is allowed to change. Unlike prior work on disentanglement and independent component analysis, we allow for both nontrivial statistical and causal dependencies in the latent space. We study the identifiability of the latent representation based on pairs of views of the observations and prove sufficient conditions that allow us to identify the invariant content partition up to an invertible mapping in both generative and discriminative settings. We find numerical simulations with dependent latent variables are consistent with our theory. Lastly, we introduce Causal3DIdent, a dataset of high-dimensional, visually complex images with rich causal dependencies, which we use to study the effect of data augmentations performed in practice.
****
Instance-Conditional Knowledge Distillation for Object Detection
Zijian Kang, Peizhen Zhang, Xiangyu Zhang, Jian Sun, Nanning Zheng
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Self-Supervised Representation Learning on Neural Network Weights for Model Characteristic Prediction
Konstantin Schürholt, Dimche Kostadinov, Damian Borth
Self-Supervised Learning (SSL) has been shown to learn useful and information-pr

eserving representations. Neural Networks (NNs) are widely applied, yet their we
ight space is still not fully understood. Therefore, we propose to use SSL to le
arn hyper-representations of the weights of populations of NNs. To that end, we
introduce domain specific data augmentations and an adapted attention architectu
re. Our empirical evaluation demonstrates that self-supervised representation l
earning in this domain is able to recover diverse NN model characteristics. Furt
her, we show that the proposed learned representations outperform prior work for
 predicting hyper-parameters, test accuracy, and generalization gap as well as t
ransfer to out-of-distribution settings.
****
Multimodal Virtual Point 3D Detection
Tianwei Yin, Xingyi Zhou, Philipp Krähenbühl
Requests for name changes in the electronic proceedings will be accepted with 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 Joint Learning for Solving Placement and Routing in Chip Design
Ruoyu Cheng, Junchi Yan
For its advantage in GPU acceleration and less dependency on human experts, mach
ine learning has been an emerging tool for solving the placement and routing pro
blems, as two critical steps in modern chip design flow. Being still in its earl
y stage, there are several fundamental issues unresolved: scalability, reward de
sign, and end-to-end learning paradigm etc. To achieve end-to-end placement lear
ning, we first propose a joint learning method for the placement of macros and s
tandard cells, by the integration of reinforcement learning with a gradient base
d optimization scheme. To further bridge the placement with the subsequent routi
ng task, we also develop a joint learning approach via reinforcement learning. O
ne key design in our (reinforcement) learning paradigm involves a multi-view emb
edding model to encode both global graph level and local node level information
of the input macros. Moreover, the random network distillation is devised to enc
ourage exploration. Experiments on public chip design benchmarks show that our m
ethod can effectively learn from experience and also provide high-quality interm
ediate placement for the post standard cell placement, within few hours for trai
ning.
****
Learning with Algorithmic Supervision via Continuous Relaxations
Felix Petersen, Christian Borgelt, Hilde Kuehne, Oliver Deussen
The integration of algorithmic components into neural architectures has gained i
ncreased attention recently, as it allows training neural networks with new form
s of supervision such as ordering constraints or silhouettes instead of using gr
ound truth labels. Many approaches in the field focus on the continuous relaxati
on of a specific task and show promising results in this context. But the focus
on single tasks also limits the applicability of the proposed concepts to a narr
ow range of applications. In this work, we build on those ideas to propose an ap
proach that allows to integrate algorithms into end-to-end trainable neural netw
ork architectures based on a general approximation of discrete conditions. To th
is end, we relax these conditions in control structures such as conditional stat
ements, loops, and indexing, so that resulting algorithms are smoothly different
iable. To obtain meaningful gradients, each relevant variable is perturbed via l
ogistic distributions and the expectation value under this perturbation is appro
ximated. We evaluate the proposed continuous relaxation model on four challengin
g tasks and show that it can keep up with relaxations specifically designed for
each individual task.
****
Differentiable Multiple Shooting Layers
Stefano Massaroli, Michael Poli, Sho Sonoda, Taiji Suzuki, Jinkyoo Park, Atsushi
 Yamashita, Hajime Asama
We detail a novel class of implicit neural models. Leveraging time-parallel meth
ods for differential equations, Multiple Shooting Layers (MSLs) seek solutions

of initial value problems via parallelizable root-finding algorithms. MSLs broadly serve as drop-in replacements for neural ordinary differential equations (Neural ODEs) with improved efficiency in number of function evaluations (NFEs) and wall-clock inference time. We develop the algorithmic framework of MSLs, analyzing the different choices of solution methods from a theoretical and computational perspective. MSLs are showcased in long horizon optimal control of ODEs and PDEs and as latent models for sequence generation. Finally, we investigate the speedups obtained through application of MSL inference in neural controlled differential equations (Neural CDEs) for time series classification of medical data.
****
Global-aware Beam Search for Neural Abstractive Summarization
Ye Ma, Zixun Lan, Lu Zong, Kaizhu Huang
This study develops a calibrated beam-based algorithm with awareness of the global attention distribution for neural abstractive summarization, aiming to improve the local optimality problem of the original beam search in a rigorous way. Specifically, a novel global protocol is proposed based on the attention distribution to stipulate how a global optimal hypothesis should attend to the source. A global scoring mechanism is then developed to regulate beam search to generate summaries in a near-global optimal fashion. This novel design enjoys a distinctive property, i.e., the global attention distribution could be predicted before inference, enabling step-wise improvements on the beam search through the global scoring mechanism. Extensive experiments on nine datasets show that the global (attention)-aware inference significantly improves state-of-the-art summarization models even using empirical hyper-parameters. The algorithm is also proven robust as it remains to generate meaningful texts with corrupted attention distributions. The codes and a comprehensive set of examples are available.
****
DROID-SLAM: Deep Visual SLAM for Monocular, Stereo, and RGB-D Cameras
Zachary Teed, Jia Deng
We introduce DROID-SLAM, a new deep learning based SLAM system. DROID-SLAM consists of recurrent iterative updates of camera pose and pixelwise depth through a Dense Bundle Adjustment layer. DROID-SLAM is accurate, achieving large improvements over prior work, and robust, suffering from substantially fewer catastrophic failures. Despite training on monocular video, it can leverage stereo or RGB-D video to achieve improved performance at test time. The URL to our open source code is https://github.com/princeton-vl/DROID-SLAM.
****
Few-Shot Object Detection via Association and DIscrimination
Yuhang Cao, Jiaqi Wang, Ying Jin, Tong Wu, Kai Chen, Ziwei Liu, Dahua Lin
Object detection has achieved substantial progress in the last decade. However, detecting novel classes with only few samples remains challenging, since deep learning under low data regime usually leads to a degraded feature space. Existing works employ a holistic fine-tuning paradigm to tackle this problem, where the model is first pre-trained on all base classes with abundant samples, and then it is used to carve the novel class feature space. Nonetheless, this paradigm is still imperfect. Durning fine-tuning, a novel class may implicitly leverage the knowledge of multiple base classes to construct its feature space, which induces a scattered feature space, hence violating the inter-class separability. To overcome these obstacles, we propose a two-step fine-tuning framework, Few-shot object detection via Association and DIscrimination (FADI), which builds up a discriminative feature space for each novel class with two integral steps. 1) In the association step, in contrast to implicitly leveraging multiple base classes, we construct a compact novel class feature space via explicitly imitating a specific base class feature space. Specifically, we associate each novel class with a base class according to their semantic similarity. After that, the feature space of a novel class can readily imitate the well-trained feature space of the associated base class. 2) In the discrimination step, to ensure the separability between the novel classes and associated base classes, we disentangle the classification branches for base and novel classes. To further enlarge the inter-class separability between all classes, a set-specialized margin loss is imposed. Extens

ive experiments on standard Pascal VOC and MS-COCO datasets demonstrate that FAD I achieves new state-of-the-art performance, significantly improving the baselin e in any shot/split by +18.7. Notably, the advantage of FADI is most announced o n extremely few-shot scenarios (e.g. 1- and 3- shot).
****

Neural Dubber: Dubbing for Videos According to Scripts

Chenxu Hu, Qiao Tian, Tingle Li, Wang Yuping, Yuxuan Wang, Hang Zhao

Dubbing is a post-production process of re-recording actors' dialogues, which is extensively used in filmmaking and video production. It is usually performed ma nually by professional voice actors who read lines with proper prosody, and in s ynchronization with the pre-recorded videos. In this work, we propose Neural Dub ber, the first neural network model to solve a novel automatic video dubbing (AV D) task: synthesizing human speech synchronized with the given video from the te xt. Neural Dubber is a multi-modal text-to-speech (TTS) model that utilizes the lip movement in the video to control the prosody of the generated speech. Furthe rmore, an image-based speaker embedding (ISE) module is developed for the multi-speaker setting, which enables Neural Dubber to generate speech with a reasonabl e timbre according to the speaker's face. Experiments on the chemistry lecture s ingle-speaker dataset and LRS2 multi-speaker dataset show that Neural Dubber can generate speech audios on par with state-of-the-art TTS models in terms of spee ch quality. Most importantly, both qualitative and quantitative evaluations show that Neural Dubber can control the prosody of synthesized speech by the video, and generate high-fidelity speech temporally synchronized with the video.
****

Neural Bootstrapper

Minsuk Shin, Hyungjoo Cho, Hyun-seok Min, Sungbin Lim

Bootstrapping has been a primary tool for ensemble and uncertainty quantificatio n in machine learning and statistics. However, due to its nature of multiple tra ining and resampling, bootstrapping deep neural networks is computationally burd ensome; hence it has difficulties in practical application to the uncertainty es timation and related tasks. To overcome this computational bottleneck, we propos e a novel approach called Neural Bootstrapper (NeuBoots), which learns to genera te bootstrapped neural networks through single model training. NeuBoots injects the bootstrap weights into the high-level feature layers of the backbone network and outputs the bootstrapped predictions of the target, without additional para meters and the repetitive computations from scratch. We apply NeuBoots to variou s machine learning tasks related to uncertainty quantification, including predic tion calibrations in image classification and semantic segmentation, active lear ning, and detection of out-of-distribution samples. Our empirical results show t hat NeuBoots outperforms other bagging based methods under a much lower computat ional cost without losing the validity of bootstrapping.
****

An Axiomatic Theory of Provably-Fair Welfare-Centric Machine Learning

Cyrus Cousins

We address an inherent difficulty in welfare-theoretic fair machine learning (ML), by proposing an equivalently-axiomatically justified alternative setting, and studying the resulting computational and statistical learning questions. Welfar e metrics quantify overall wellbeing across a population of groups, and welfare-based objectives and constraints have recently been proposed to incentivize fair ML methods to satisfy their diverse needs. However, many ML problems are cast a s loss minimization tasks, rather than utility maximization, and thus require no ntrivial modeling to construct utility functions. We define a complementary metr ic, termed malfare, measuring overall societal harm, with axiomatic justificatio n via the standard axioms of cardinal welfare, and cast fair ML as malfare minim ization over the risk values (expected losses) of each group. Surprisingly, the axioms of cardinal welfare (malfare) dictate that this is not equivalent to simp ly defining utility as negative loss and maximizing welfare. Building upon these concepts, we define fair-PAC learning, where a fair-PAC learner is an algorithm that learns an ε-δ malfare-optimal model with bounded sample complexity, for an y data distribution and (axiomatically justified) malfare concept. Finally, we s

how conditions under which many standard PAC-learners may be converted to fair-PAC learners, which places fair-PAC learning on firm theoretical ground, as it yields statistical — and in some cases computational — efficiency guarantees for many well-studied ML models. Fair-PAC learning is also practically relevant, as it democratizes fair ML by providing concrete training algorithms with rigorous generalization guarantees.
****
HSVA: Hierarchical Semantic-Visual Adaptation for Zero-Shot Learning
Shiming Chen, Guosen Xie, Yang Liu, Qinmu Peng, Baigui Sun, Hao Li, Xinge You, Ling Shao
Zero-shot learning (ZSL) tackles the unseen class recognition problem, transferring semantic knowledge from seen classes to unseen ones. Typically, to guarantee desirable knowledge transfer, a common (latent) space is adopted for associating the visual and semantic domains in ZSL. However, existing common space learning methods align the semantic and visual domains by merely mitigating distribution disagreement through one-step adaptation. This strategy is usually ineffective due to the heterogeneous nature of the feature representations in the two domains, which intrinsically contain both distribution and structure variations. To address this and advance ZSL, we propose a novel hierarchical semantic-visual adaptation (HSVA) framework. Specifically, HSVA aligns the semantic and visual domains by adopting a hierarchical two-step adaptation, i.e., structure adaptation and distribution adaptation. In the structure adaptation step, we take two task-specific encoders to encode the source data (visual domain) and the target data (semantic domain) into a structure-aligned common space. To this end, a supervised adversarial discrepancy (SAD) module is proposed to adversarially minimize the discrepancy between the predictions of two task-specific classifiers, thus making the visual and semantic feature manifolds more closely aligned. In the distribution adaptation step, we directly minimize the Wasserstein distance between the latent multivariate Gaussian distributions to align the visual and semantic distributions using a common encoder. Finally, the structure and distribution adaptation are derived in a unified framework under two partially-aligned variational autoencoders. Extensive experiments on four benchmark datasets demonstrate that HSVA achieves superior performance on both conventional and generalized ZSL. The code is available at \url{https://github.com/shiming-chen/HSVA}.
****
Higher Order Kernel Mean Embeddings to Capture Filtrations of Stochastic Processes
Cristopher Salvi, Maud Lemercier, Chong Liu, Blanka Horvath, Theodoros Damoulas, Terry Lyons
Stochastic processes are random variables with values in some space of paths. However, reducing a stochastic process to a path-valued random variable ignores its filtration, i.e. the flow of information carried by the process through time. By conditioning the process on its filtration, we introduce a family of higher order kernel mean embeddings (KMEs) that generalizes the notion of KME to capture additional information related to the filtration. We derive empirical estimators for the associated higher order maximum mean discrepancies (MMDs) and prove consistency. We then construct a filtration-sensitive kernel two-sample test able to capture information that gets missed by the standard MMD test. In addition, leveraging our higher order MMDs we construct a family of universal kernels on stochastic processes that allows to solve real-world calibration and optimal stopping problems in quantitative finance (such as the pricing of American options) via classical kernel-based regression methods. Finally, adapting existing tests for conditional independence to the case of stochastic processes, we design a causal-discovery algorithm to recover the causal graph of structural dependencies among interacting bodies solely from observations of their multidimensional trajectories.
****
Low-Rank Subspaces in GANs
Jiapeng Zhu, Ruili Feng, Yujun Shen, Deli Zhao, Zheng-Jun Zha, Jingren Zhou, Qifeng Chen

The latent space of a Generative Adversarial Network (GAN) has been shown to encode rich semantics within some subspaces. To identify these subspaces, researchers typically analyze the statistical information from a collection of synthesized data, and the identified subspaces tend to control image attributes globally (i.e., manipulating an attribute causes the change of an entire image). By contrast, this work introduces low-rank subspaces that enable more precise control of GAN generation. Concretely, given an arbitrary image and a region of interest (e.g., eyes of face images), we manage to relate the latent space to the image region with the Jacobian matrix and then use low-rank factorization to discover steerable latent subspaces. There are three distinguishable strengths of our approach that can be aptly called LowRankGAN. First, compared to analytic algorithms in prior work, our low-rank factorization of Jacobians is able to find the low-dimensional representation of attribute manifold, making image editing more precise and controllable. Second, low-rank factorization naturally yields a null space of attributes such that moving the latent code within it only affects the outer region of interest. Therefore, local image editing can be simply achieved by projecting an attribute vector into the null space without relying on a spatial mask as existing methods do. Third, our method can robustly work with a local region from one image for analysis yet well generalize to other images, making it much easy to use in practice. Extensive experiments on state-of-the-art GAN models (including StyleGAN2 and BigGAN) trained on various datasets demonstrate the effectiveness of our LowRankGAN.
****
Neural Symplectic Form: Learning Hamiltonian Equations on General Coordinate Systems
Yuhan Chen, Takashi Matsubara, Takaharu Yaguchi
In recent years, substantial research on the methods for learning Hamiltonian equations has been conducted. Although these approaches are very promising, the commonly used representation of the Hamilton equation uses the generalized momenta, which are generally unknown. Therefore, the training data must be represented in this unknown coordinate system, and this causes difficulty in applying the model to real data. Meanwhile, Hamiltonian equations also have a coordinate-free expression that is expressed by using the symplectic 2-form. In this study, we propose a model that learns the symplectic form from data using neural networks, thereby providing a method for learning Hamiltonian equations from data represented in general coordinate systems, which are not limited to the generalized coordinates and the generalized momenta. Consequently, the proposed method is capable not only of modeling target equations of both Hamiltonian and Lagrangian formalisms but also of extracting unknown Hamiltonian structures hidden in the data. For example, many polynomial ordinary differential equations such as the Lotka-Volterra equation are known to admit non-trivial Hamiltonian structures, and our numerical experiments show that such structures can be certainly learned from data. Technically, each symplectic 2-form is associated with a skew-symmetric matrix, but not all skew-symmetric matrices define the symplectic 2-form. In the proposed method, using the fact that symplectic 2-forms are derived as the exterior derivative of certain differential 1-forms, we model the differential 1-form by neural networks, thereby improving the efficiency of learning.
****
Sample-Efficient Reinforcement Learning Is Feasible for Linearly Realizable MDPs with Limited Revisiting
Gen Li, Yuxin Chen, Yuejie Chi, Yuantao Gu, 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.
****
Self-Paced Contrastive Learning for Semi-supervised Medical Image Segmentation with Meta-labels
Jizong Peng, Ping Wang, Christian Desrosiers, Marco Pedersoli
The contrastive pre-training of a recognition model on a large dataset of unlabe

led data often boosts the model’s performance on downstream tasks like image classification. However, in domains such as medical imaging, collecting unlabeled data can be challenging and expensive. In this work, we consider the task of medical image segmentation and adapt contrastive learning with meta-label annotations to scenarios where no additional unlabeled data is available. Meta-labels, such as the location of a 2D slice in a 3D MRI scan, often come for free during the acquisition process. We use these meta-labels to pre-train the image encoder, as well as in a semi-supervised learning step that leverages a reduced set of annotated data. A self-paced learning strategy exploiting the weak annotations is proposed to furtherhelp the learning process and discriminate useful labels from noise. Results on five medical image segmentation datasets show that our approach: i) highly boosts the performance of a model trained on a few scans, ii) outperforms previous contrastive and semi-supervised approaches, and iii) reaches close to the performance of a model trained on the full data.
****
Reverse engineering recurrent neural networks with Jacobian switching linear dynamical systems
Jimmy Smith, Scott Linderman, David Sussillo
Recurrent neural networks (RNNs) are powerful models for processing time-series data, but it remains challenging to understand how they function. Improving this understanding is of substantial interest to both the machine learning and neuroscience communities. The framework of reverse engineering a trained RNN by linearizing around its fixed points has provided insight, but the approach has significant challenges. These include difficulty choosing which fixed point to expand around when studying RNN dynamics and error accumulation when reconstructing the nonlinear dynamics with the linearized dynamics. We present a new model that overcomes these limitations by co-training an RNN with a novel switching linear dynamical system (SLDS) formulation. A first-order Taylor series expansion of the co-trained RNN and an auxiliary function trained to pick out the RNN's fixed points govern the SLDS dynamics. The results are a trained SLDS variant that closely approximates the RNN, an auxiliary function that can produce a fixed point for each point in state-space, and a trained nonlinear RNN whose dynamics have been regularized such that its first-order terms perform the computation, if possible. This model removes the post-training fixed point optimization and allows us to unambiguously study the learned dynamics of the SLDS at any point in state-space. It also generalizes SLDS models to continuous manifolds of switching points while sharing parameters across switches. We validate the utility of the model on two synthetic tasks relevant to previous work reverse engineering RNNs. We then show that our model can be used as a drop-in in more complex architectures, such as LFADS, and apply this LFADS hybrid to analyze single-trial spiking activity from the motor system of a non-human primate.
****
Learning-Augmented Dynamic Power Management with Multiple States via New Ski Rental Bounds
Antonios Antoniadis, Christian Coester, Marek Elias, Adam Polak, Bertrand Simon
We study the online problem of minimizing power consumption in systems with multiple power-saving states. During idle periods of unknown lengths, an algorithm has to choose between power-saving states of different energy consumption and wake-up costs. We develop a learning-augmented online algorithm that makes decisions based on (potentially inaccurate) predicted lengths of the idle periods. The algorithm's performance is near-optimal when predictions are accurate and degrades gracefully with increasing prediction error, with a worst-case guarantee almost identical to the optimal classical online algorithm for the problem. A key ingredient in our approach is a new algorithm for the online ski-rental problem in the learning augmented setting with tight dependence on the prediction error. We support our theoretical findings with experiments.
****
Learning Equivariant Energy Based Models with Equivariant Stein Variational Gradient Descent
Priyank Jaini, Lars Holdijk, Max Welling

We focus on the problem of efficient sampling and learning of probability densit
ies by incorporating symmetries in probabilistic models. We first introduce Equi
variant Stein Variational Gradient Descent algorithm -- an equivariant sampling
method based on Stein's identity for sampling from densities with symmetries. Eq
uivariant SVGD explicitly incorporates symmetry information in a density through
 equivariant kernels which makes the resultant sampler efficient both in terms o
f sample complexity and the quality of generated samples. Subsequently, we defin
e equivariant energy based models to model invariant densities that are learned
using contrastive divergence. By utilizing our equivariant SVGD for training equ
ivariant EBMs, we propose new ways of improving and scaling up training of energ
y based models. We apply these equivariant energy models for modelling joint den
sities in regression and classification tasks for image datasets, many-body part
icle systems and molecular structure generation.
****

Information Directed Sampling for Sparse Linear Bandits
Botao Hao, Tor Lattimore, Wei Deng
Stochastic sparse linear bandits offer a practical model for high-dimensional on
line decision-making problems and have a rich information-regret structure. In t
his work we explore the use of information-directed sampling (IDS), which natura
lly balances the information-regret trade-off. We develop a class of information
-theoretic Bayesian regret bounds that nearly match existing lower bounds on a v
ariety of problem instances, demonstrating the adaptivity of IDS. To efficiently
 implement sparse IDS, we propose an empirical Bayesian approach for sparse post
erior sampling using a spike-and-slab Gaussian-Laplace prior. Numerical results
 demonstrate significant regret reductions by sparse IDS relative to several bas
elines.
****

Linear Convergence of Gradient Methods for Estimating Structured Transition Matr
ices in High-dimensional Vector Autoregressive Models
Xiao Lv, Wei Cui, Yulong Liu
In this paper, we present non-asymptotic optimization guarantees of gradient des
cent methods for estimating structured transition matrices in high-dimensional v
ector autoregressive (VAR) models. We adopt the projected gradient descent (PGD)
 for single-structured transition matrices and the alternating projected gradien
t descent (AltPGD) for superposition-structured ones. Our analysis demonstrates
that both gradient algorithms converge linearly to the statistical error even th
ough the strong convexity of the objective function is absent under the high-dim
ensional settings. Moreover our result is sharp (up to a constant factor) in the
 sense of matching the phase transition theory of the corresponding model with
independent samples. To the best of our knowledge, this analysis constitutes fir
st non-asymptotic optimization guarantees of the linear rate for regularized est
imation in high-dimensional VAR models. Numerical results are provided to suppo
rt our theoretical analysis.
****

Large-Scale Unsupervised Object Discovery
Van Huy Vo, Elena Sizikova, Cordelia Schmid, Patrick Pérez, Jean Ponce
Existing approaches to unsupervised object discovery (UOD) do not scale up to la
rge datasets without approximations that compromise their performance. We propos
e a novel formulation of UOD as a ranking problem, amenable to the arsenal of di
stributed methods available for eigenvalue problems and link analysis. Through t
he use of self-supervised features, we also demonstrate the first effective ful
ly unsupervised pipeline for UOD. Extensive experiments on COCO~\cite{Lin2014coc
odataset} and OpenImages~\cite{openimages} show that, in the single-object disco
very setting where a single prominent object is sought in each image, the propos
ed LOD (Large-scale Object Discovery) approach is on par with, or better than th
e state of the art for medium-scale datasets (up to 120K images), and over 37\%
better than the only other algorithms capable of scaling up to 1.7M images. In t
he multi-object discovery setting where multiple objects are sought in each imag
e, the proposed LOD is over 14\% better in average precision (AP) than all other
 methods for datasets ranging from 20K to 1.7M images. Using self-supervised fea

tures, we also show that the proposed method obtains state-of-the-art UOD perfor mance on OpenImages.
****
Sparse Steerable Convolutions: An Efficient Learning of SE(3)-Equivariant Featur es for Estimation and Tracking of Object Poses in 3D Space
Jiehong Lin, Hongyang Li, Ke Chen, Jiangbo Lu, Kui Jia
As a basic component of SE(3)-equivariant deep feature learning, steerable convo lution has recently demonstrated its advantages for 3D semantic analysis. The ad vantages are, however, brought by expensive computations on dense, volumetric da ta, which prevent its practical use for efficient processing of 3D data that are inherently sparse. In this paper, we propose a novel design of Sparse Steerable Convolution (SS-Conv) to address the shortcoming; SS-Conv greatly accelerates s teerable convolution with sparse tensors, while strictly preserving the property of SE(3)-equivariance. Based on SS-Conv, we propose a general pipeline for prec ise estimation of object poses, wherein a key design is a Feature-Steering modul e that takes the full advantage of SE(3)-equivariance and is able to conduct an efficient pose refinement. To verify our designs, we conduct thorough experimen ts on three tasks of 3D object semantic analysis, including instance-level 6D po se estimation, category-level 6D pose and size estimation, and category-level 6D pose tracking. Our proposed pipeline based on SS-Conv outperforms existing meth ods on almost all the metrics evaluated by the three tasks. Ablation studies als o show the superiority of our SS-Conv over alternative convolutions in terms of both accuracy and efficiency. Our code is released publicly at https://github.co m/Gorilla-Lab-SCUT/SS-Conv.
****
Noisy Adaptation Generates Lévy Flights in Attractor Neural Networks
Xingsi Dong, Tianhao Chu, Tiejun Huang, Zilong Ji, Si Wu
Lévy flights describe a special class of random walks whose step sizes satisfy a power-law tailed distribution. As being an efficientsearching strategy in unkno wn environments, Lévy flights are widely observed in animal foraging behaviors. Recent studies further showed that human cognitive functions also exhibit the ch aracteristics of Lévy flights. Despite being a general phenomenon, the neural m echanism at the circuit level for generating Lévy flights remains unresolved. He re, we investigate how Lévy flights can be achieved in attractor neural networks . To elucidate the underlying mechanism clearly, we first study continuous attra ctor neural networks (CANNs), and find that noisy neural adaptation, exemplified by spike frequency adaptation (SFA) in this work, can generate Lévy flights re presenting transitions of the network state in the attractor space. Specifically , the strength of SFA defines a travelling wave boundary, below which the networ k state displays local Brownian motion, and above which the network state displa ys long-jump motion. Noises in neural adaptation causes the network state to int ermittently switch between these two motion modes, manifesting the characteristi cs of Lévy flights. We further extend the study to a general attractor neural ne twork, and demonstrate that our model can explain the Lévy-flight phenomenon obs erved during free memory retrieval of humans. We hope that this study will give us insight into understanding the neural mechanism for optimal information proce ssing in the brain.
****
On Linear Stability of SGD and Input-Smoothness of Neural Networks
Chao Ma, Lexing Ying
The multiplicative structure of parameters and input data in the first layer of neural networks is explored to build connection between the landscape of the los s function with respect to parameters and the landscape of the model function wi th respect to input data. By this connection, it is shown that flat minima regul arize the gradient of the model function, which explains the good generalization performance of flat minima. Then, we go beyond the flatness and consider high-o rder moments of the gradient noise, and show that Stochastic Gradient Dascent (S GD) tends to impose constraints on these moments by a linear stability analysis of SGD around global minima. Together with the multiplicative structure, we iden tify the Sobolev regularization effect of SGD, i.e. SGD regularizes the Sobolev

seminorms of the model function with respect to the input data. Finally, bounds for generalization error and adversarial robustness are provided for solutions found by SGD under assumptions of the data distribution.
****
Joint inference and input optimization in equilibrium networks
Swaminathan Gurumurthy, Shaojie Bai, Zachary Manchester, J. Zico Kolter
Many tasks in deep learning involve optimizing over the inputs to a network to minimize or maximize some objective; examples include optimization over latent spaces in a generative model to match a target image, or adversarially perturbing an input to worsen classifier performance. Performing such optimization, however, is traditionally quite costly, as it involves a complete forward and backward pass through the network for each gradient step. In a separate line of work, a recent thread of research has developed the deep equilibrium (DEQ) model, a class of models that foregoes traditional network depth and instead computes the output of a network by finding the fixed point of a single nonlinear layer. In this paper, we show that there is a natural synergy between these two settings. Although, naively using DEQs for these optimization problems is expensive (owing to the time needed to compute a fixed point for each gradient step), we can leverage the fact that gradient-based optimization can itself be cast as a fixed point iteration to substantially improve the overall speed. That is, we simultaneously both solve for the DEQ fixed point and optimize over network inputs, all within a single "augmented" DEQ model that jointly encodes both the original network and the optimization process. Indeed, the procedure is fast enough that it allows us to efficiently train DEQ models for tasks traditionally relying on an "inner" optimization loop. We demonstrate this strategy on various tasks such as training generative models while optimizing over latent codes, training models for inverse problems like denoising and inpainting, adversarial training and gradient based meta-learning.
****
A unified framework for bandit multiple testing
Ziyu Xu, Ruodu Wang, Aaditya Ramdas
In bandit multiple hypothesis testing, each arm corresponds to a different null hypothesis that we wish to test, and the goal is to design adaptive algorithms that correctly identify large set of interesting arms (true discoveries), while only mistakenly identifying a few uninteresting ones (false discoveries). One common metric in non-bandit multiple testing is the false discovery rate (FDR). We propose a unified, modular framework for bandit FDR control that emphasizes the decoupling of exploration and summarization of evidence. We utilize the powerful martingale-based concept of "e-processes" to ensure FDR control for arbitrary composite nulls, exploration rules and stopping times in generic problem settings. In particular, valid FDR control holds even if the reward distributions of the arms could be dependent, multiple arms may be queried simultaneously, and multiple (cooperating or competing) agents may be querying arms, covering combinatorial semi-bandit type settings as well. Prior work has considered in great detail the setting where each arm's reward distribution is independent and sub-Gaussian, and a single arm is queried at each step. Our framework recovers matching sample complexity guarantees in this special case, and performs comparably or better in practice. For other settings, sample complexities will depend on the finer details of the problem (composite nulls being tested, exploration algorithm, data dependence structure, stopping rule) and we do not explore these; our contribution is to show that the FDR guarantee is clean and entirely agnostic to these details.
****
Recovering Latent Causal Factor for Generalization to Distributional Shifts
Xinwei Sun, Botong Wu, Xiangyu Zheng, Chang Liu, Wei Chen, Tao Qin, Tie-Yan Liu
Distributional shifts between training and target domains may degrade the prediction accuracy of learned models, mainly because these models often learn features that possess only correlation rather than causal relation with the output. Such a correlation, which is known as ``spurious correlation'' statistically, is domain-dependent hence may fail to generalize to unseen domains. To avoid such a s

purious correlation, we propose \textbf{La}tent \textbf{C}ausal \textbf{I}nvaria nce \textbf{M}odels (LaCIM) that specifies the underlying causal structure of th e data and the source of distributional shifts, guiding us to pursue only causal factor for prediction. Specifically, the LaCIM introduces a pair of correlated latent factors: (a) causal factor and (b) others, while the extent of this corre lation is governed by a domain variable that characterizes the distributional sh ifts. On the basis of this, we prove that the distribution of observed variables conditioning on latent variables is shift-invariant. Equipped with such an inva riance, we prove that the causal factor can be recovered without mixing informat ion from others, which induces the ground-truth predicting mechanism. We propose a Variational-Bayesian-based method to learn this invariance for prediction. Th e utility of our approach is verified by improved generalization to distribution al shifts on various real-world data. Our code is freely available at \url{https ://github.com/wubotong/LaCIM}.
****
Graph Differentiable Architecture Search with Structure Learning
Yijian Qin, Xin Wang, Zeyang Zhang, Wenwu Zhu
Discovering ideal Graph Neural Networks (GNNs) architectures for different tasks is labor intensive and time consuming. To save human efforts, Neural Architectu re Search (NAS) recently has been used to automatically discover adequate GNN ar chitectures for certain tasks in order to achieve competitive or even better per formance compared with manually designed architectures. However, existing works utilizing NAS to search GNN structures fail to answer the question: how NAS is a ble to select the desired GNN architectures? In this paper, we investigate this question to solve the problem, for the first time. We conduct a measurement stud y with experiments to discover that gradient based NAS methods tend to select pr oper architectures based on the usefulness of different types of information wit h respect to the target task. Our explorations further show that gradient based NAS also suffers from noises hidden in the graph, resulting in searching subopti mal GNN architectures. Based on our findings, we propose a Graph differentiable Architecture Search model with Structure Optimization (GASSO), which allows diff erentiable search of the architecture with gradient descent and is able to disco ver graph neural architectures with better performance through employing graph s tructure learning as a denoising process in the search procedure. The proposed G ASSO model is capable of simultaneously searching the optimal architecture and a daptively adjusting graph structure by jointly optimizing graph architecture sea rch and graph structure denoising. Extensive experiments on real-world graph dat asets demonstrate that our proposed GASSO model is able to achieve state-of-the-art performance compared with existing baselines.
****
Designing Counterfactual Generators using Deep Model Inversion
Jayaraman Thiagarajan, Vivek Sivaraman Narayanaswamy, Deepta Rajan, Jia Liang, A kshay Chaudhari, Andreas Spanias
Explanation techniques that synthesize small, interpretable changes to a given i mage while producing desired changes in the model prediction have become popular for introspecting black-box models. Commonly referred to as counterfactuals, th e synthesized explanations are required to contain discernible changes (for easy interpretability) while also being realistic (consistency to the data manifold) . In this paper, we focus on the case where we have access only to the trained d eep classifier and not the actual training data. While the problem of inverting deep models to synthesize images from the training distribution has been explore d, our goal is to develop a deep inversion approach to generate counterfactual e xplanations for a given query image. Despite their effectiveness in conditional image synthesis, we show that existing deep inversion methods are insufficient f or producing meaningful counterfactuals. We propose DISC (Deep Inversion for Syn thesizing Counterfactuals) that improves upon deep inversion by utilizing (a) st ronger image priors, (b) incorporating a novel manifold consistency objective an d (c) adopting a progressive optimization strategy. We find that, in addition to producing visually meaningful explanations, the counterfactuals from DISC are e ffective at learning classifier decision boundaries and are robust to unknown te

st-time corruptions.
****
A Faster Maximum Cardinality Matching Algorithm with Applications in Machine Lea
rning
Nathaniel Lahn, Sharath Raghvendra, Jiacheng 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.
****
Dynamic population-based meta-learning for multi-agent communication with natura
l language
Abhinav Gupta, Marc Lanctot, Angeliki Lazaridou
In this work, our goal is to train agents that can coordinate with seen, unseen
as well as human partners in a multi-agent communication environment involving n
atural language. Previous work using a single set of agents has shown great prog
ress in generalizing to known partners, however it struggles when coordinating w
ith unfamiliar agents. To mitigate that, recent work explored the use of populat
ion-based approaches, where multiple agents interact with each other with the go
al of learning more generic protocols. These methods, while able to result in go
od coordination between unseen partners, still only achieve so in cases of simpl
e languages, thus failing to adapt to human partners using natural language. We
attribute this to the use of static populations and instead propose a dynamic po
pulation-based meta-learning approach that builds such a population in an iterat
ive manner. We perform a holistic evaluation of our method on two different refe
rential games, and show that our agents outperform all prior work when communica
ting with seen partners and humans. Furthermore, we analyze the natural language
 generation skills of our agents, where we find that our agents also outperform
strong baselines. Finally, we test the robustness of our agents when communicati
ng with out-of-population agents and carefully test the importance of each compo
nent of our method through ablation studies.
****
Adversarial Neuron Pruning Purifies Backdoored Deep Models
Dongxian Wu, Yisen Wang
As deep neural networks (DNNs) are growing larger, their requirements for comput
ational resources become huge, which makes outsourcing training more popular. Tr
aining in a third-party platform, however, may introduce potential risks that a
malicious trainer will return backdoored DNNs, which behave normally on clean sa
mples but output targeted misclassifications whenever a trigger appears at the t
est time. Without any knowledge of the trigger, it is difficult to distinguish o
r recover benign DNNs from backdoored ones. In this paper, we first identify an
unexpected sensitivity of backdoored DNNs, that is, they are much easier to coll
apse and tend to predict the target label on clean samples when their neurons ar
e adversarially perturbed. Based on these observations, we propose a novel model
 repairing method, termed Adversarial Neuron Pruning (ANP), which prunes some se
nsitive neurons to purify the injected backdoor. Experiments show, even with onl
y an extremely small amount of clean data (e.g., 1%), ANP effectively removes th
e injected backdoor without causing obvious performance degradation.
****
Towards Robust and Reliable Algorithmic Recourse
Sohini Upadhyay, Shalmali Joshi, Himabindu Lakkaraju
As predictive models are increasingly being deployed in high-stakes decision mak
ing (e.g., loan approvals), there has been growing interest in post-hoc techniqu
es which provide recourse to affected individuals. These techniques generate re
courses under the assumption that the underlying predictive model does not chang
e. However, in practice, models are often regularly updated for a variety of rea
sons (e.g., dataset shifts), thereby rendering previously prescribed recourses i
neffective.To address this problem, we propose a novel framework, RObust Algorit
hmic Recourse (ROAR), that leverages adversarial training for finding recourses
that are robust to model shifts. To the best of our knowledge, this work propose

s the first ever solution to this critical problem. We also carry out theoretical analysis which underscores the importance of constructing recourses that are robust to model shifts: 1) We quantify the probability of invalidation for recourses generated without accounting for model shifts. 2) We prove that the additional cost incurred due to the robust recourses output by our framework is bounded. Experimental evaluation on multiple synthetic and real-world datasets demonstrates the efficacy of the proposed framework.
****
Neural Rule-Execution Tracking Machine For Transformer-Based Text Generation
Yufei Wang, Can Xu, Huang Hu, Chongyang Tao, Stephen Wan, Mark Dras, Mark Johnson, Daxin Jiang
Sequence-to-Sequence (Seq2Seq) neural text generation models, especially the pre-trained ones (e.g., BART and T5), have exhibited compelling performance on various natural language generation tasks. However, the black-box nature of these models limits their application in tasks where specific rules (e.g., controllable constraints, prior knowledge) need to be executed. Previous works either design specific model structures (e.g., Copy Mechanism corresponding to the rule "the generated output should include certain words in the source input'') or implement specialized inference algorithms (e.g., Constrained Beam Search) to execute particular rules through the text generation. These methods require the careful design case-by-case and are difficult to support multiple rules concurrently. In this paper, we propose a novel module named Neural Rule-Execution Tracking Machine (NRETM) that can be equipped into various transformer-based generators to leverage multiple rules simultaneously to guide the neural generation model for superior generation performance in an unified and scalable way. Extensive experiments on several benchmarks verify the effectiveness of our proposed model in both controllable and general text generation tasks.
****
Scalable Online Planning via Reinforcement Learning Fine-Tuning
Arnaud Fickinger, Hengyuan Hu, Brandon Amos, Stuart Russell, Noam Brown
Lookahead search has been a critical component of recent AI successes, such as in the games of chess, go, and poker. However, the search methods used in these games, and in many other settings, are tabular. Tabular search methods do not scale well with the size of the search space, and this problem is exacerbated by stochasticity and partial observability. In this work we replace tabular search with online model-based fine-tuning of a policy neural network via reinforcement learning, and show that this approach outperforms state-of-the-art search algorithms in benchmark settings. In particular, we use our search algorithm to achieve a new state-of-the-art result in self-play Hanabi, and show the generality of our algorithm by also showing that it outperforms tabular search in the Atari game Ms. Pacman.
****
Adversarial Regression with Doubly Non-negative Weighting Matrices
Tam Le, Truyen Nguyen, Makoto Yamada, Jose Blanchet, Viet Anh Nguyen
Many machine learning tasks that involve predicting an output response can be solved by training a weighted regression model. Unfortunately, the predictive power of this type of models may severely deteriorate under low sample sizes or under covariate perturbations. Reweighting the training samples has aroused as an effective mitigation strategy to these problems. In this paper, we propose a novel and coherent scheme for kernel-reweighted regression by reparametrizing the sample weights using a doubly non-negative matrix. When the weighting matrix is confined in an uncertainty set using either the log-determinant divergence or the Bures-Wasserstein distance, we show that the adversarially reweighted estimate can be solved efficiently using first-order methods. Numerical experiments show that our reweighting strategy delivers promising results on numerous datasets.
****
Learned Robust PCA: A Scalable Deep Unfolding Approach for High-Dimensional Outlier Detection
HanQin Cai, Jialin Liu, Wotao Yin
Robust principal component analysis (RPCA) is a critical tool in modern machine

learning, which detects outliers in the task of low-rank matrix reconstruction. In this paper, we propose a scalable and learnable non-convex approach for high-dimensional RPCA problems, which we call Learned Robust PCA (LRPCA). LRPCA is highly efficient, and its free parameters can be effectively learned to optimize via deep unfolding. Moreover, we extend deep unfolding from finite iterations to infinite iterations via a novel feedforward-recurrent-mixed neural network model. We establish the recovery guarantee of LRPCA under mild assumptions for RPCA. Numerical experiments show that LRPCA outperforms the state-of-the-art RPCA algorithms, such as ScaledGD and AltProj, on both synthetic datasets and real-world applications.
****
Proxy-Normalizing Activations to Match Batch Normalization while Removing Batch Dependence
Antoine Labatie, Dominic Masters, Zach Eaton-Rosen, Carlo Luschi
We investigate the reasons for the performance degradation incurred with batch-independent normalization. We find that the prototypical techniques of layer normalization and instance normalization both induce the appearance of failure modes in the neural network's pre-activations: (i) layer normalization induces a collapse towards channel-wise constant functions; (ii) instance normalization induces a lack of variability in instance statistics, symptomatic of an alteration of the expressivity. To alleviate failure mode (i) without aggravating failure mode (ii), we introduce the technique "Proxy Normalization" that normalizes post-activations using a proxy distribution. When combined with layer normalization or group normalization, this batch-independent normalization emulates batch normalization's behavior and consistently matches or exceeds its performance.
****
Dynamic Bottleneck for Robust Self-Supervised Exploration
Chenjia Bai, Lingxiao Wang, Lei Han, Animesh Garg, Jianye Hao, Peng Liu, Zhaoran Wang
Exploration methods based on pseudo-count of transitions or curiosity of dynamics have achieved promising results in solving reinforcement learning with sparse rewards. However, such methods are usually sensitive to environmental dynamics-irrelevant information, e.g., white-noise. To handle such dynamics-irrelevant information, we propose a Dynamic Bottleneck (DB) model, which attains a dynamics-relevant representation based on the information-bottleneck principle. Based on the DB model, we further propose DB-bonus, which encourages the agent to explore state-action pairs with high information gain. We establish theoretical connections between the proposed DB-bonus, the upper confidence bound (UCB) for linear case, and the visiting count for tabular case. We evaluate the proposed method on Atari suits with dynamics-irrelevant noises. Our experiments show that exploration with DB bonus outperforms several state-of-the-art exploration methods in noisy environments.
****
ProTo: Program-Guided Transformer for Program-Guided Tasks
Zelin Zhao, Karan Samel, Binghong Chen, lee song
Programs, consisting of semantic and structural information, play an important role in the communication between humans and agents. Towards learning general program executors to unify perception, reasoning, and decision making, we formulate program-guided tasks which require learning to execute a given program on the observed task specification. Furthermore, we propose Program-Guided Transformers (ProTo), which integrates both semantic and structural guidance of a program by leveraging cross-attention and masked self-attention to pass messages between the specification and routines in the program. ProTo executes a program in a learned latent space and enjoys stronger representation ability than previous neural-symbolic approaches. We demonstrate that ProTo significantly outperforms the previous state-of-the-art methods on GQA visual reasoning and 2D Minecraft policy learning datasets. Additionally, ProTo demonstrates better generalization to unseen, complex, and human-written programs.
****
An Efficient Transfer Learning Framework for Multiagent Reinforcement Learning

Tianpei Yang, Weixun Wang, Hongyao Tang, Jianye Hao, Zhaopeng Meng, Hangyu Mao, Dong Li, Wulong Liu, Yingfeng Chen, Yujing Hu, Changjie Fan, Chengwei Zhang
Transfer Learning has shown great potential to enhance single-agent Reinforcement Learning (RL) efficiency. Similarly, Multiagent RL (MARL) can also be accelerated if agents can share knowledge with each other. However, it remains a problem of how an agent should learn from other agents. In this paper, we propose a novel Multiagent Policy Transfer Framework (MAPTF) to improve MARL efficiency. MAPTF learns which agent's policy is the best to reuse for each agent and when to terminate it by modeling multiagent policy transfer as the option learning problem. Furthermore, in practice, the option module can only collect all agent's local experiences for update due to the partial observability of the environment. While in this setting, each agent's experience may be inconsistent with each other, which may cause the inaccuracy and oscillation of the option-value's estimation. Therefore, we propose a novel option learning algorithm, the successor representation option learning to solve it by decoupling the environment dynamics from rewards and learning the option-value under each agent's preference. MAPTF can be easily combined with existing deep RL and MARL approaches, and experimental results show it significantly boosts the performance of existing methods in both discrete and continuous state spaces.
****
Learning to Time-Decode in Spiking Neural Networks Through the Information Bottleneck
Nicolas Skatchkovsky, Osvaldo Simeone, Hyeryung Jang
One of the key challenges in training Spiking Neural Networks (SNNs) is that target outputs typically come in the form of natural signals, such as labels for classification or images for generative models, and need to be encoded into spikes. This is done by handcrafting target spiking signals, which in turn implicitly fixes the mechanisms used to decode spikes into natural signals, e.g., rate decoding. The arbitrary choice of target signals and decoding rule generally impairs the capacity of the SNN to encode and process information in the timing of spikes. To address this problem, this work introduces a hybrid variational autoencoder architecture, consisting of an encoding SNN and a decoding Artificial Neural Network (ANN). The role of the decoding ANN is to learn how to best convert the spiking signals output by the SNN into the target natural signal. A novel end-to-end learning rule is introduced that optimizes a directed information bottleneck training criterion via surrogate gradients. We demonstrate the applicability of the technique in an experimental settings on various tasks, including real-life datasets.
****
NEO: Non Equilibrium Sampling on the Orbits of a Deterministic Transform
Achille Thin, Yazid Janati El Idrissi, Sylvain Le Corff, Charles Ollion, Eric Moulines, Arnaud Doucet, Alain Durmus, Christian X Robert
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Relaxing Local Robustness
Klas Leino, Matt Fredrikson
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Tuning Large Neural Networks via Zero-Shot Hyperparameter Transfer
Ge Yang, Edward Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, Jianfeng Gao
Requests for name changes in the electronic proceedings will be accepted with 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.
****
Statistical Regeneration Guarantees of the Wasserstein Autoencoder with Latent Space Consistency
Anish Chakrabarty, Swagatam Das
The introduction of Variational Autoencoders (VAE) has been marked as a breakthrough in the history of representation learning models. Besides having several accolades of its own, VAE has successfully flagged off a series of inventions in the form of its immediate successors. Wasserstein Autoencoder (WAE), being an heir to that realm carries with it all of the goodness and heightened generative promises, matching even the generative adversarial networks (GANs). Needless to say, recent years have witnessed a remarkable resurgence in statistical analyses of the GANs. Similar examinations for Autoencoders however, despite their diverse applicability and notable empirical performance, remain largely absent. To close this gap, in this paper, we investigate the statistical properties of WAE. Firstly, we provide statistical guarantees that WAE achieves the target distribution in the latent space, utilizing the Vapnik–Chervonenkis (VC) theory. The main result, consequently ensures the regeneration of the input distribution, harnessing the potential offered by Optimal Transport of measures under the Wasserstein metric. This study, in turn, hints at the class of distributions WAE can reconstruct after suffering a compression in the form of a latent law.
****
Leveraging the Inductive Bias of Large Language Models for Abstract Textual Reasoning
Christopher Rytting, David Wingate
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Differentiable Simulation of Soft Multi-body Systems
Yiling Qiao, Junbang Liang, Vladlen Koltun, Ming Lin
We present a method for differentiable simulation of soft articulated bodies. Our work enables the integration of differentiable physical dynamics into gradient-based pipelines. We develop a top-down matrix assembly algorithm within Projective Dynamics and derive a generalized dry friction model for soft continuum using a new matrix splitting strategy. We derive a differentiable control framework for soft articulated bodies driven by muscles, joint torques, or pneumatic tubes. The experiments demonstrate that our designs make soft body simulation more stable and realistic compared to other frameworks. Our method accelerates the solution of system identification problems by more than an order of magnitude, and enables efficient gradient-based learning of motion control with soft robots.
****
Good Classification Measures and How to Find Them
Martijn Gösgens, Anton Zhiyanov, Aleksey Tikhonov, Liudmila Prokhorenkova
Several performance measures can be used for evaluating classification results: accuracy, F-measure, and many others. Can we say that some of them are better than others, or, ideally, choose one measure that is best in all situations? To answer this question, we conduct a systematic analysis of classification performance measures: we formally define a list of desirable properties and theoretically analyze which measures satisfy which properties. We also prove an impossibility theorem: some desirable properties cannot be simultaneously satisfied. Finally, we propose a new family of measures satisfying all desirable properties except one. This family includes the Matthews Correlation Coefficient and a so-called Symmetric Balanced Accuracy that was not previously used in classification literature. We believe that our systematic approach gives an important tool to practitioners for adequately evaluating classification results.
****
Distilling Robust and Non-Robust Features in Adversarial Examples by Information Bottleneck

Junho Kim, Byung-Kwan Lee, Yong Man Ro
Adversarial examples, generated by carefully crafted perturbation, have attracte
d considerable attention in research fields. Recent works have argued that the e
xistence of the robust and non-robust features is a primary cause of the adversa
rial examples, and investigated their internal interactions in the feature space
. In this paper, we propose a way of explicitly distilling feature representatio
n into the robust and non-robust features, using Information Bottleneck. Specifi
cally, we inject noise variation to each feature unit and evaluate the informati
on flow in the feature representation to dichotomize feature units either robust
 or non-robust, based on the noise variation magnitude. Through comprehensive ex
periments, we demonstrate that the distilled features are highly correlated with
 adversarial prediction, and they have human-perceptible semantic information by
 themselves. Furthermore, we present an attack mechanism intensifying the gradie
nt of non-robust features that is directly related to the model prediction, and
validate its effectiveness of breaking model robustness.
****
Vector-valued Gaussian Processes on Riemannian Manifolds via Gauge Independent P
rojected Kernels
Michael Hutchinson, Alexander Terenin, Viacheslav Borovitskiy, So Takao, Yee Teh
, Marc Deisenroth
Gaussian processes are machine learning models capable of learning unknown funct
ions in a way that represents uncertainty, thereby facilitating construction of
optimal decision-making systems. Motivated by a desire to deploy Gaussian proces
ses in novel areas of science, a rapidly-growing line of research has focused on
 constructively extending these models to handle non-Euclidean domains, includin
g Riemannian manifolds, such as spheres and tori. We propose techniques that gen
eralize this class to model vector fields on Riemannian manifolds, which are imp
ortant in a number of application areas in the physical sciences. To do so, we p
resent a general recipe for constructing gauge independent kernels, which induce
 Gaussian vector fields, i.e. vector-valued Gaussian processes coherent withgeom
etry, from scalar-valued Riemannian kernels. We extend standard Gaussian process
 training methods, such as variational inference, to this setting. This enables
vector-valued Gaussian processes on Riemannian manifolds to be trained using sta
ndard methods and makes them accessible to machine learning practitioners.
****
On the Representation Power of Set Pooling Networks
Christian Bueno, Alan Hylton
Point clouds and sets are input data-types which pose unique problems to deep le
arning. Since sets can have variable cardinality and are unchanged by permutatio
n, the input space for these problems naturally form infinite-dimensional non-Eu
clidean spaces. Despite these mathematical difficulties, PointNet (Qi et al. 201
7) and Deep Sets (Zaheer et al. 2017) introduced foundational neural network arc
hitectures to address these problems. In this paper we present a unified framewo
rk to study the expressive power of such networks as well as their extensions be
yond point clouds (partially addressing a conjecture on the extendibility of Dee
pSets along the way). To this end, we demonstrate the crucial role that the Haus
dorff and Wasserstein metrics play and prove new cardinality-agnostic universali
ty results to characterize exactly which functions can be approximated by these
models. In particular, these results imply that PointNet generally cannot approx
imate averages of continuous functions over sets (e.g. center-of-mass or higher
moments) implying that DeepSets is strictly more expressive than PointNet in the
 constant cardinality setting. Moreover, we obtain explicit lower-bounds on the
approximation error and present a simple method to produce arbitrarily many exam
ples of this failure-mode. Counterintuitively, we also prove that in the unbound
ed cardinality setting that any function which can be uniformly approximated by
both PointNet and normalized-DeepSets must be constant. Finally, we also prove t
heorems on the Lipschitz properties of PointNet and normalized-DeepSets which sh
ed insight into exploitable inductive bias in these networks.
****
Learning Policies with Zero or Bounded Constraint Violation for Constrained MDPs

Tao Liu, Ruida Zhou, Dileep Kalathil, Panganamala Kumar, Chao 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.
****
A Prototype-Oriented Framework for Unsupervised Domain Adaptation
Korawat Tanwisuth, Xinjie Fan, Huangjie Zheng, Shujian Zhang, Hao Zhang, Bo Chen, Mingyuan Zhou
Existing methods for unsupervised domain adaptation often rely on minimizing some statistical distance between the source and target samples in the latent space. To avoid the sampling variability, class imbalance, and data-privacy concerns that often plague these methods, we instead provide a memory and computation-efficient probabilistic framework to extract class prototypes and align the target features with them. We demonstrate the general applicability of our method on a wide range of scenarios, including single-source, multi-source, class-imbalance, and source-private domain adaptation. Requiring no additional model parameters and having a moderate increase in computation over the source model alone, the proposed method achieves competitive performance with state-of-the-art methods.
****
Mining the Benefits of Two-stage and One-stage HOI Detection
Aixi Zhang, Yue Liao, Si Liu, Miao Lu, Yongliang Wang, Chen Gao, XIAOBO LI
Two-stage methods have dominated Human-Object Interaction~(HOI) detection for several years. Recently, one-stage HOI detection methods have become popular. In this paper, we aim to explore the essential pros and cons of two-stage and one-stage methods. With this as the goal, we find that conventional two-stage methods mainly suffer from positioning positive interactive human-object pairs, while one-stage methods are challenging to make an appropriate trade-off on multi-task learning, \emph{i.e.}, object detection, and interaction classification. Therefore, a core problem is how to take the essence and discard the dregs from the conventional two types of methods. To this end, we propose a novel one-stage framework with disentangling human-object detection and interaction classification in a cascade manner. In detail, we first design a human-object pair generator based on a state-of-the-art one-stage HOI detector by removing the interaction classification module or head and then design a relatively isolated interaction classifier to classify each human-object pair. Two cascade decoders in our proposed framework can focus on one specific task, detection or interaction classification. In terms of the specific implementation, we adopt a transformer-based HOI detector as our base model. The newly introduced disentangling paradigm outperforms existing methods by a large margin, with a significant relative mAP gain of 9.32% on HICO-Det. The source codes are available at https://github.com/YueLiao/CDN.
****
Discerning Decision-Making Process of Deep Neural Networks with Hierarchical Voting Transformation
Ying Sun, Hengshu Zhu, Chuan Qin, Fuzhen Zhuang, Qing He, Hui Xiong
Neural network based deep learning techniques have shown great success for numerous applications. While it is expected to understand their intrinsic decision-making processes, these deep neural networks often work in a black-box way. To this end, in this paper, we aim to discern the decision-making processes of neural networks through a hierarchical voting strategy by developing an explainable deep learning model, namely Voting Transformation-based Explainable Neural Network (VOTEN). Specifically, instead of relying on massive feature combinations, VOTEN creatively models expressive single-valued voting functions between explicitly modeled latent concepts to achieve high fitting ability. Along this line, we first theoretically analyze the major components of VOTEN and prove the relationship and advantages of VOTEN compared with Multi-Layer Perceptron (MLP), the basic structure of deep neural networks. Moreover, we design efficient algorithms to improve the model usability by explicitly showing the decision processes of VOTEN. Finally, extensive experiments on multiple real-world datasets clearly validate the performances and explainability of VOTEN.

****
Risk-averse Heteroscedastic Bayesian Optimization
Anastasia Makarova, Ilnura Usmanova, Ilija Bogunovic, Andreas Krause
Many black-box optimization tasks arising in high-stakes applications require risk-averse decisions. The standard Bayesian optimization (BO) paradigm, however, optimizes the expected value only. We generalize BO to trade mean and input-dependent variance of the objective, both of which we assume to be unknown a priori. In particular, we propose a novel risk-averse heteroscedastic Bayesian optimization algorithm (RAHBO) that aims to identify a solution with high return and low noise variance, while learning the noise distribution on the fly. To this end, we model both expectation and variance as (unknown) RKHS functions, and propose a novel risk-aware acquisition function. We bound the regret for our approach and provide a robust rule to report the final decision point for applications where only a single solution must be identified. We demonstrate the effectiveness of RAHBO on synthetic benchmark functions and hyperparameter tuning tasks.
****
Invertible DenseNets with Concatenated LipSwish
Yura Perugachi-Diaz, Jakub Tomczak, Sandjai Bhulai
We introduce Invertible Dense Networks (i-DenseNets), a more parameter efficient extension of Residual Flows. The method relies on an analysis of the Lipschitz continuity of the concatenation in DenseNets, where we enforce invertibility of the network by satisfying the Lipschitz constant. Furthermore, we propose a learnable weighted concatenation, which not only improves the model performance but also indicates the importance of the concatenated weighted representation. Additionally, we introduce the Concatenated LipSwish as activation function, for which we show how to enforce the Lipschitz condition and which boosts performance. The new architecture, i-DenseNet, out-performs Residual Flow and other flow-based models on density estimation evaluated in bits per dimension, where we utilize an equal parameter budget. Moreover, we show that the proposed model out-performs Residual Flows when trained as a hybrid model where the model is both a generative and a discriminative model.
****
Topological Detection of Trojaned Neural Networks
Songzhu Zheng, Yikai Zhang, Hubert Wagner, Mayank Goswami, Chao Chen
Deep neural networks are known to have security issues. One particular threat is the Trojan attack. It occurs when the attackers stealthily manipulate the model's behavior through Trojaned training samples, which can later be exploited. Guided by basic neuroscientific principles, we discover subtle -- yet critical -- structural deviation characterizing Trojaned models. In our analysis we use topological tools. They allow us to model high-order dependencies in the networks, robustly compare different networks, and localize structural abnormalities. One interesting observation is that Trojaned models develop short-cuts from shallow to deep layers. Inspired by these observations, we devise a strategy for robust detection of Trojaned models. Compared to standard baselines it displays better performance on multiple benchmarks.
****
Provably Strict Generalisation Benefit for Invariance in Kernel Methods
Bryn Elesedy
It is a commonly held belief that enforcing invariance improves generalisation. Although this approach enjoys widespread popularity, it is only very recently that a rigorous theoretical demonstration of this benefit has been established. In this work we build on the function space perspective of Elesedy and Zaidi [8] to derive a strictly non-zero generalisation benefit of incorporating invariance in kernel ridge regression when the target is invariant to the action of a compact group. We study invariance enforced by feature averaging and find that generalisation is governed by a notion of effective dimension that arises from the interplay between the kernel and the group. In building towards this result, we find that the action of the group induces an orthogonal decomposition of both the reproducing kernel Hilbert space and its kernel, which may be of interest in its own right.

****
Formalizing the Generalization-Forgetting Trade-off in Continual Learning
Krishnan Raghavan, Prasanna Balaprakash
We formulate the continual learning (CL) problem via dynamic programming and model the trade-off between catastrophic forgetting and generalization as a two-player sequential game. In this approach, player 1 maximizes the cost due to lack of generalization whereas player 2 minimizes the cost due to catastrophic forgetting. We show theoretically that a balance point between the two players exists for each task and that this point is stable (once the balance is achieved, the two players stay at the balance point). Next, we introduce balanced continual learning (BCL), which is designed to attain balance between generalization and forgetting and empirically demonstrate that BCL is comparable to or better than the state of the art.
****
Risk-Aware Transfer in Reinforcement Learning using Successor Features
Michael Gimelfarb, Andre Barreto, Scott Sanner, Chi-Guhn Lee
Sample efficiency and risk-awareness are central to the development of practical reinforcement learning (RL) for complex decision-making. The former can be addressed by transfer learning, while the latter by optimizing some utility function of the return. However, the problem of transferring skills in a risk-aware manner is not well-understood. In this paper, we address the problem of transferring policies between tasks in a common domain that differ only in their reward functions, in which risk is measured by the variance of reward streams. Our approach begins by extending the idea of generalized policy improvement to maximize entropic utilities, thus extending the dynamic programming's policy improvement operation to sets of policies \emph{and} levels of risk-aversion. Next, we extend the idea of successor features (SF), a value function representation that decouples the environment dynamics from the rewards, to capture the variance of returns. Our resulting risk-aware successor features (RaSF) integrate seamlessly within the RL framework, inherit the superior task generalization ability of SFs, while incorporating risk into the decision-making. Experiments on a discrete navigation domain and control of a simulated robotic arm demonstrate the ability of RaSFs to outperform alternative methods including SFs, when taking the risk of the learned policies into account.
****
Causal Inference for Event Pairs in Multivariate Point Processes
Tian Gao, Dharmashankar Subramanian, Debarun Bhattacharjya, Xiao Shou, Nicholas Mattei, Kristin P Bennett
Causal inference and discovery from observational data has been extensively studied across multiple fields. However, most prior work has focused on independent and identically distributed (i.i.d.) data. In this paper, we propose a formalization for causal inference between pairs of event variables in multivariate recurrent event streams by extending Rubin's framework for the average treatment effect (ATE) and propensity scores to multivariate point processes. Analogous to a joint probability distribution representing i.i.d. data, a multivariate point process represents data involving asynchronous and irregularly spaced occurrences of various types of events over a common timeline. We theoretically justify our point process causal framework and show how to obtain unbiased estimates of the proposed measure. We conduct an experimental investigation using synthetic and real-world event datasets, where our proposed causal inference framework is shown to exhibit superior performance against a set of baseline pairwise causal association scores.
****
Evaluating model performance under worst-case subpopulations
Mike Li, Hongseok Namkoong, Shangzhou Xia
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

Privately Publishable Per-instance Privacy
Rachel Redberg, Yu-Xiang Wang
We consider how to privately share the personalized privacy losses incurred by objective perturbation, using per-instance differential privacy (pDP). Standard differential privacy (DP) gives us a worst-case bound that might be orders of magnitude larger than the privacy loss to a particular individual relative to a fixed dataset. The pDP framework provides a more fine-grained analysis of the privacy guarantee to a target individual, but the per-instance privacy loss itself might be a function of sensitive data. In this paper, we analyze the per-instance privacy loss of releasing a private empirical risk minimizer learned via objective perturbation, and propose a group of methods to privately and accurately publish the pDP losses at little to no additional privacy cost.
****
Understanding the Limits of Unsupervised Domain Adaptation via Data Poisoning
Akshay Mehra, Bhavya Kailkhura, Pin-Yu Chen, Jihun Hamm
Unsupervised domain adaptation (UDA) enables cross-domain learning without target domain labels by transferring knowledge from a labeled source domain whose distribution differs from that of the target. However, UDA is not always successful and several accounts of `negative transfer' have been reported in the literature. In this work, we prove a simple lower bound on the target domain error that complements the existing upper bound. Our bound shows the insufficiency of minimizing source domain error and marginal distribution mismatch for a guaranteed reduction in the target domain error, due to the possible increase of induced labeling function mismatch. This insufficiency is further illustrated through simple distributions for which the same UDA approach succeeds, fails, and may succeed or fail with an equal chance. Motivated from this, we propose novel data poisoning attacks to fool UDA methods into learning representations that produce large target domain errors. We evaluate the effect of these attacks on popular UDA methods using benchmark datasets where they have been previously shown to be successful. Our results show that poisoning can significantly decrease the target domain accuracy, dropping it to almost 0% in some cases, with the addition of only 10% poisoned data in the source domain. The failure of these UDA methods demonstrates their limitations at guaranteeing cross-domain generalization consistent with our lower bound. Thus, evaluating UDA methods in adversarial settings such as data poisoning provides a better sense of their robustness to data distributions unfavorable for UDA.
****
Coresets for Clustering with Missing Values
Vladimir Braverman, Shaofeng Jiang, Robert Krauthgamer, Xuan 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.
****
Boosting with Multiple Sources
Corinna Cortes, Mehryar Mohri, Dmitry Storcheus, Ananda Theertha Suresh
We study the problem of learning accurate ensemble predictors, in particular boosting, in the presence of multiple source domains. We show that the standard convex combination ensembles in general cannot succeed in this scenario and adopt instead a domain-weighted combination. We introduce and analyze a new boosting algorithm, MULTIBOOST, for this scenario and show that it benefits from favorable theoretical guarantees. We also report the results of several experiments with our algorithm demonstrating that it outperforms natural baselines on multi-source text-based, image-based and tabular data. We further present an extension of our algorithm to the federated learning scenario and report favorable experimental results for that setting as well. Additionally, we describe in detail an extension of our algorithm to the multi-class setting, MCMULTIBOOST, for which we also report experimental results.
****
Dynamic Neural Representational Decoders for High-Resolution Semantic Segmentati

on
Bowen Zhang, Yifan liu, Zhi Tian, Chunhua Shen
Requests for name changes in the electronic proceedings will be accepted with no
 questions asked. However name changes may cause bibliographic tracking issues.
  Authors are asked to consider this carefully and discuss it with their co-auth
ors prior to requesting a name change in the electronic proceedings.
****
Dense Keypoints via Multiview Supervision
Zhixuan Yu, Haozheng Yu, Long Sha, Sujoy Ganguly, Hyun Soo Park
This paper presents a new end-to-end semi-supervised framework to learn a dense
keypoint detector using unlabeled multiview images. A key challenge lies in ■ndi
ng the exact correspondences between the dense keypoints in multiple views since
 the inverse of the keypoint mapping can be neither analytically derived nor dif
ferentiated. This limits applying existing multiview supervision approaches used
 to learn sparse keypoints that rely on the exact correspondences. To address th
is challenge, we derive a new probabilistic epipolar constraint that encodes the
 two desired properties. (1) Soft correspondence: we de■ne a matchability, which
 measures a likelihood of a point matching to the other image’s corresponding po
int, thus relaxing the requirement of the exact correspondences. (2) Geometric c
onsistency: every point in the continuous correspondence ■elds must satisfy the
multiview consistency collectively. We formulate a probabilistic epipolar constr
aint using a weighted average of epipolar errors through the matchability thereb
y generalizing the point-to-point geometric error to the ■eld-to-■eld geometric
error. This generalization facilitates learning a geometrically coherent dense k
eypoint detection model by utilizing a large number of unlabeled multiview image
s. Additionally, to prevent degenerative cases, we employ a distillation-based r
egularization by using a pretrained model. Finally, we design a new neural netwo
rk architecture, made of twin networks, that effectively minimizes the probabili
stic epipolar errors of all possible correspondences between two view images by
building af■nity matrices. Our method shows superior performance compared to exi
sting methods, including non-differentiable bootstrapping in terms of keypoint a
ccuracy, multiview consistency, and 3D reconstruction accuracy.
****
Scatterbrain: Unifying Sparse and Low-rank Attention
Beidi Chen, Tri Dao, Eric Winsor, Zhao Song, Atri Rudra, Christopher Ré
Requests for name changes in the electronic proceedings will be accepted with 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.
****
PTR: A Benchmark for Part-based Conceptual, Relational, and Physical Reasoning
Yining Hong, Li Yi, Josh Tenenbaum, Antonio Torralba, Chuang Gan
A critical aspect of human visual perception is the ability to parse visual scen
es into individual objects and further into object parts, forming part-whole hie
rarchies. Such composite structures could induce a rich set of semantic concepts
 and relations, thus playing an important role in the interpretation and organiz
ation of visual signals as well as for the generalization of visual perception a
nd reasoning. However, existing visual reasoning benchmarks mostly focus on obje
cts rather than parts. Visual reasoning based on the full part-whole hierarchy i
s much more challenging than object-centric reasoning due to finer-grained conce
pts, richer geometry relations, and more complex physics. Therefore, to better s
erve for part-based conceptual, relational and physical reasoning, we introduce
a new large-scale diagnostic visual reasoning dataset named PTR. PTR contains ar
ound 80k RGBD synthetic images with ground truth object and part level annotatio
ns regarding semantic instance segmentation, color attributes, spatial and geome
tric relationships, and certain physical properties such as stability. These ima
ges are paired with 800k machine-generated questions covering various types of r
easoning types, making them a good testbed for visual reasoning models. We exami
ne several state-of-the-art visual reasoning models on this dataset and observe
that they still make many surprising mistakes in situations where humans can eas

ily infer the correct answer. We believe this dataset will open up new opportunities for part-based reasoning. PTR dataset and baseline models are publicly available.
****
Property-Aware Relation Networks for Few-Shot Molecular Property Prediction
Yaqing Wang, Abulikemu Abuduweili, Quanming Yao, Dejing Dou
Molecular property prediction plays a fundamental role in drug discovery to identify candidate molecules with target properties. However, molecular property prediction is essentially a few-shot problem, which makes it hard to use regular machine learning models. In this paper, we propose Property-Aware Relation networks (PAR) to handle this problem. In comparison to existing works, we leverage the fact that both relevant substructures and relationships among molecules change across different molecular properties. We first introduce a property-aware embedding function to transform the generic molecular embeddings to substructure-aware space relevant to the target property. Further, we design an adaptive relation graph learning module to jointly estimate molecular relation graph and refine molecular embeddings w.r.t. the target property, such that the limited labels can be effectively propagated among similar molecules. We adopt a meta-learning strategy where the parameters are selectively updated within tasks in order to model generic and property-aware knowledge separately. Extensive experiments on benchmark molecular property prediction datasets show that PAR consistently outperforms existing methods and can obtain property-aware molecular embeddings and model molecular relation graph properly.
****
Differentially Private Learning with Adaptive Clipping
Galen Andrew, Om Thakkar, Brendan McMahan, Swaroop Ramaswamy
Existing approaches for training neural networks with user-level differential privacy (e.g., DP Federated Averaging) in federated learning (FL) settings involve bounding the contribution of each user's model update by {\em clipping} it to some constant value. However there is no good {\em a priori} setting of the clipping norm across tasks and learning settings: the update norm distribution depends on the model architecture and loss, the amount of data on each device, the client learning rate, and possibly various other parameters. We propose a method wherein instead of a fixed clipping norm, one clips to a value at a specified quantile of the update norm distribution, where the value at the quantile is itself estimated online, with differential privacy. The method tracks the quantile closely, uses a negligible amount of privacy budget, is compatible with other federated learning technologies such as compression and secure aggregation, and has a straightforward joint DP analysis with DP-FedAvg. Experiments demonstrate that adaptive clipping to the median update norm works well across a range of federated learning tasks, eliminating the need to tune any clipping hyperparameter.
****
Can Less be More? When Increasing-to-Balancing Label Noise Rates Considered Beneficial
Yang Liu, Jialu Wang
In this paper, we answer the question of when inserting label noise (less informative labels) can instead return us more accurate and fair models. We are primarily inspired by three observations: 1) In contrast to reducing label noise rates, increasing the noise rates is easy to implement; 2) Increasing a certain class of instances' label noise to balance the noise rates (increasing-to-balancing) results in an easier learning problem; 3) Increasing-to-balancing improves fairness guarantees against label bias. In this paper, we first quantify the trade-offs introduced by increasing a certain group of instances' label noise rate w.r.t. the loss of label informativeness and the lowered learning difficulties. We analytically demonstrate when such an increase is beneficial, in terms of either improved generalization power or the fairness guarantees. Then we present a method to insert label noise properly for the task of learning with noisy labels, either without or with a fairness constraint. The primary technical challenge we face is due to the fact that we would not know which data instances are suffering from higher noise, and we would not have the ground truth labels to verify any p

ossible hypothesis. We propose a detection method that informs us which group of labels might suffer from higher noise without using ground truth labels. We formally establish the effectiveness of the proposed solution and demonstrate it with extensive experiments.
****

Projected GANs Converge Faster

Axel Sauer, Kashyap Chitta, Jens Müller, Andreas Geiger

Generative Adversarial Networks (GANs) produce high-quality images but are challenging to train. They need careful regularization, vast amounts of compute, and expensive hyper-parameter sweeps. We make significant headway on these issues by projecting generated and real samples into a fixed, pretrained feature space. Motivated by the finding that the discriminator cannot fully exploit features from deeper layers of the pretrained model, we propose a more effective strategy that mixes features across channels and resolutions. Our Projected GAN improves image quality, sample efficiency, and convergence speed. It is further compatible with resolutions of up to one Megapixel and advances the state-of-the-art Fréchet Inception Distance (FID) on twenty-two benchmark datasets. Importantly, Projected GANs match the previously lowest FIDs up to 40 times faster, cutting the wall-clock time from 5 days to less than 3 hours given the same computational resources.
****

Generating High-Quality Explanations for Navigation in Partially-Revealed Environments

Gregory Stein

We present an approach for generating natural language explanations of high-level behavior of autonomous agents navigating in partially-revealed environments. Our counterfactual explanations communicate changes to interpratable statistics of the belief (e.g., the likelihood an exploratory action will reach the unseen goal) that are estimated from visual input via a deep neural network and used (via a Bellman equation variant) to inform planning far into the future. Additionally, our novel training procedure mimics explanation generation, allowing us to use planning performance as an objective measure of explanation quality. Simulated experiments validate that our explanations are both high quality and can be used in interventions to directly correct bad behavior; agents trained via our training-by-explaining procedure achieve 9.1% lower average cost than a non-learned baseline (12.7% after interventions) in environments derived from real-world floor plans.
****

De-randomizing MCMC dynamics with the diffusion Stein operator

Zheyang Shen, Markus Heinonen, Samuel Kaski

Approximate Bayesian inference estimates descriptors of an intractable target distribution - in essence, an optimization problem within a family of distributions. For example, Langevin dynamics (LD) extracts asymptotically exact samples from a diffusion process because the time evolution of its marginal distributions constitutes a curve that minimizes the KL-divergence via steepest descent in the Wasserstein space. Parallel to LD, Stein variational gradient descent (SVGD) similarly minimizes the KL, albeit endowed with a novel Stein-Wasserstein distance, by deterministically transporting a set of particle samples, thus de-randomizes the stochastic diffusion process. We propose de-randomized kernel-based particle samplers to all diffusion-based samplers known as MCMC dynamics. Following previous work in interpreting MCMC dynamics, we equip the Stein-Wasserstein space with a fiber-Riemannian Poisson structure, with the capacity of characterizing a fiber-gradient Hamiltonian flow that simulates MCMC dynamics. Such dynamics discretizes into generalized SVGD (GSVGD), a Stein-type deterministic particle sampler, with particle updates coinciding with applying the diffusion Stein operator to a kernel function. We demonstrate empirically that GSVGD can de-randomize complex MCMC dynamics, which combine the advantages of auxiliary momentum variables and Riemannian structure, while maintaining the high sample quality from an interacting particle system.
****

Sparsely Changing Latent States for Prediction and Planning in Partially Observable Domains
Christian Gumbsch, Martin V. Butz, Georg Martius
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
PreferenceNet: Encoding Human Preferences in Auction Design with Deep Learning
Neehar Peri, Michael Curry, Samuel Dooley, John Dickerson
The design of optimal auctions is a problem of interest in economics, game theory and computer science. Despite decades of effort, strategyproof, revenue-maximizing auction designs are still not known outside of restricted settings. However, recent methods using deep learning have shown some success in approximating optimal auctions, recovering several known solutions and outperforming strong baselines when optimal auctions are not known. In addition to maximizing revenue, auction mechanisms may also seek to encourage socially desirable constraints such as allocation fairness or diversity. However, these philosophical notions neither have standardization nor do they have widely accepted formal definitions. In this paper, we propose PreferenceNet, an extension of existing neural-network-based auction mechanisms to encode constraints using (potentially human-provided) exemplars of desirable allocations. In addition, we introduce a new metric to evaluate an auction allocations' adherence to such socially desirable constraints and demonstrate that our proposed method is competitive with current state-of-the-art neural-network based auction designs. We validate our approach through human subject research and show that we are able to effectively capture real human preferences.
****
Large-Scale Learning with Fourier Features and Tensor Decompositions
Frederiek Wesel, Kim Batselier
Random Fourier features provide a way to tackle large-scale machine learning problems with kernel methods. Their slow Monte Carlo convergence rate has motivated the research of deterministic Fourier features whose approximation error can decrease exponentially in the number of basis functions. However, due to their tensor product extension to multiple dimensions, these methods suffer heavily from the curse of dimensionality, limiting their applicability to one, two or three-dimensional scenarios. In our approach we overcome said curse of dimensionality by exploiting the tensor product structure of deterministic Fourier features, which enables us to represent the model parameters as a low-rank tensor decomposition. We derive a monotonically converging block coordinate descent algorithm with linear complexity in both the sample size and the dimensionality of the inputs for a regularized squared loss function, allowing to learn a parsimonious model in decomposed form using deterministic Fourier features.We demonstrate by means of numerical experiments how our low-rank tensor approach obtains the same performance of the corresponding nonparametric model, consistently outperforming random Fourier features.
****
Hash Layers For Large Sparse Models
Stephen Roller, Sainbayar Sukhbaatar, arthur szlam, Jason Weston
We investigate the training of sparse layers that use different parameters for different inputs based on hashing in large Transformer models. Specifically, we modify the feedforward layer to hash to different sets of weights depending on the current token, over all tokens in the sequence. We show that this procedure either outperforms or is competitive with learning-to-route mixture-of-expert methods such as Switch Transformers and BASE Layers, while requiring no routing parameters or extra terms in the objective function such as a load balancing loss, and no sophisticated assignment algorithm. We study the performance of different hashing techniques, hash sizes and input features, and show that balanced and random hashes focused on the most local features work best, compared to either learning clusters or using longer-range context. We show our approach works wel

l both on large language modeling and dialogue tasks, and on downstream fine-tuning tasks.
****
Sliced Mutual Information: A Scalable Measure of Statistical Dependence
Ziv Goldfeld, Kristjan Greenewald
Mutual information (MI) is a fundamental measure of statistical dependence, with a myriad of applications to information theory, statistics, and machine learning. While it possesses many desirable structural properties, the estimation of high-dimensional MI from samples suffers from the curse of dimensionality. Motivated by statistical scalability to high dimensions, this paper proposes sliced MI (SMI) as a surrogate measure of dependence. SMI is defined as an average of MI terms between one-dimensional random projections. We show that it preserves many of the structural properties of classic MI, while gaining scalable computation and efficient estimation from samples. Furthermore, and in contrast to classic MI, SMI can grow as a result of deterministic transformations. This enables leveraging SMI for feature extraction by optimizing it over processing functions of raw data to identify useful representations thereof. Our theory is supported by numerical studies of independence testing and feature extraction, which demonstrate the potential gains SMI offers over classic MI for high-dimensional inference.
****
Emergent Communication under Varying Sizes and Connectivities
Jooyeon Kim, Alice Oh
Recent advances in deep neural networks allowed artificial agents to derive their own emergent languages that promote interaction, coordination, and collaboration within a group. Just as we humans have succeeded in creating a shared language that allows us to interact within a large group, can the emergent communication within an artificial group converge to a shared, agreed language? This research provides an analytical study of the shared emergent language within the group communication settings of different sizes and connectivities. As the group size increases up to hundreds, agents start to speak dissimilar languages, but the rate at which they successfully communicate is maintained. We observe the emergence of different dialects when we restrict the group communication to have local connectivities only. Finally, we provide optimization results of group communication graphs when the number of agents one can communicate with is restricted or when we penalize communication between distant agent pairs. The optimized communication graphs show superior communication success rates compared to graphs with same number of links as well as the emergence of hub nodes and scale-free networks.
****
Deep Bandits Show-Off: Simple and Efficient Exploration with Deep Networks
Rong Zhu, Mattia Rigotti
Designing efficient exploration is central to Reinforcement Learning due to the fundamental problem posed by the exploration-exploitation dilemma. Bayesian exploration strategies like Thompson Sampling resolve this trade-off in a principled way by modeling and updating the distribution of the parameters of the action-value function, the outcome model of the environment.However, this technique becomes infeasible for complex environments due to the computational intractability of maintaining probability distributions over parameters of outcome models of corresponding complexity.Moreover, the approximation techniques introduced to mitigate this issue typically result in poor exploration-exploitation trade-offs, as observed in the case of deep neural network models with approximate posterior methods that have been shown to underperform in the deep bandit scenario.In this paper we introduce Sample Average Uncertainty (SAU), a simple and efficient uncertainty measure for contextual bandits.While Bayesian approaches like Thompson Sampling estimate outcomes uncertainty indirectly by first quantifying the variability over the parameters of the outcome model, SAU is a frequentist approach that directly estimates the uncertainty of the outcomes based on the value predictions.Importantly, we show theoretically that the uncertainty measure estimated by SAU asymptotically matches the uncertainty provided by Thompson Sampling, as well as its regret bounds.Because of its simplicity SAU can be seamlessly applie

d to deep contextual bandits as a very scalable drop-in replacement for epsilon-greedy exploration.We confirm empirically our theory by showing that SAU-based exploration outperforms current state-of-the-art deep Bayesian bandit methods on several real-world datasets at modest computation cost, and make the code to reproduce our results available at \url{https://github.com/ibm/sau-explore}.
****
Regret Minimization Experience Replay in Off-Policy Reinforcement Learning
Xu-Hui Liu, Zhenghai Xue, Jingcheng Pang, Shengyi Jiang, Feng Xu, Yang Yu
In reinforcement learning, experience replay stores past samples for further reuse. Prioritized sampling is a promising technique to better utilize these samples. Previous criteria of prioritization include TD error, recentness and corrective feedback, which are mostly heuristically designed. In this work, we start from the regret minimization objective, and obtain an optimal prioritization strategy for Bellman update that can directly maximize the return of the policy. The theory suggests that data with higher hindsight TD error, better on-policiness and more accurate Q value should be assigned with higher weights during sampling. Thus most previous criteria only consider this strategy partially. We not only provide theoretical justifications for previous criteria, but also propose two new methods to compute the prioritization weight, namely ReMERN and ReMERT. ReMERN learns an error network, while ReMERT exploits the temporal ordering of states. Both methods outperform previous prioritized sampling algorithms in challenging RL benchmarks, including MuJoCo, Atari and Meta-World.
****
Relative Uncertainty Learning for Facial Expression Recognition
Yuhang Zhang, Chengrui Wang, Weihong Deng
In facial expression recognition (FER), the uncertainties introduced by inherent noises like ambiguous facial expressions and inconsistent labels raise concerns about the credibility of recognition results. To quantify these uncertainties and achieve good performance under noisy data, we regard uncertainty as a relative concept and propose an innovative uncertainty learning method called Relative Uncertainty Learning (RUL). Rather than assuming Gaussian uncertainty distributions for all datasets, RUL builds an extra branch to learn uncertainty from the relative difficulty of samples by feature mixup. Specifically, we use uncertainties as weights to mix facial features and design an add-up loss to encourage uncertainty learning. It is easy to implement and adds little or no extra computation overhead. Extensive experiments show that RUL outperforms state-of-the-art FER uncertainty learning methods in both real-world and synthetic noisy FER datasets. Besides, RUL also works well on other datasets such as CIFAR and Tiny ImageNet. The code is available at https://github.com/zyh-uaiaaaa/Relative-Uncertainty-Learning.
****
An Information-theoretic Approach to Distribution Shifts
Marco Federici, Ryota Tomioka, Patrick Forré
Safely deploying machine learning models to the real world is often a challenging process. For example, models trained with data obtained from a specific geographic location tend to fail when queried with data obtained elsewhere, agents trained in a simulation can struggle to adapt when deployed in the real world or novel environments, and neural networks that are fit to a subset of the population might carry some selection bias into their decision process.In this work, we describe the problem of data shift from an information-theoretic perspective by (i) identifying and describing the different sources of error, (ii) comparing some of the most promising objectives explored in the recent domain generalization and fair classification literature. From our theoretical analysis and empirical evaluation, we conclude that the model selection procedure needs to be guided by careful considerations regarding the observed data, the factors used for correction, and the structure of the data-generating process.
****
TRS: Transferability Reduced Ensemble via Promoting Gradient Diversity and Model Smoothness
Zhuolin Yang, Linyi Li, Xiaojun Xu, Shiliang Zuo, Qian Chen, Pan Zhou, Benjamin

Rubinstein, Ce Zhang, Bo Li
Adversarial Transferability is an intriguing property - adversarial perturbation crafted against one model is also effective against another model, while these models are from different model families or training processes. To better protect ML systems against adversarial attacks, several questions are raised: what are the sufficient conditions for adversarial transferability, and how to bound it? Is there a way to reduce the adversarial transferability in order to improve the robustness of an ensemble ML model? To answer these questions, in this work we first theoretically analyze and outline sufficient conditions for adversarial transferability between models; then propose a practical algorithm to reduce the transferability between base models within an ensemble to improve its robustness. Our theoretical analysis shows that only promoting the orthogonality between gradients of base models is not enough to ensure low transferability; in the meantime, the model smoothness is an important factor to control the transferability. We also provide the lower and upper bounds of adversarial transferability under certain conditions. Inspired by our theoretical analysis, we propose an effective Transferability Reduced Smooth (TRS) ensemble training strategy to train a robust ensemble with low transferability by enforcing both gradient orthogonality and model smoothness between base models. We conduct extensive experiments on TRS and compare with 6 state-of-the-art ensemble baselines against 8 whitebox attacks on different datasets, demonstrating that the proposed TRS outperforms all baselines significantly.
****
Towards Sample-Optimal Compressive Phase Retrieval with Sparse and Generative Priors
Zhaoqiang Liu, Subhroshekhar Ghosh, 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.
****
Moser Flow: Divergence-based Generative Modeling on Manifolds
Noam Rozen, Aditya Grover, Maximilian Nickel, Yaron Lipman
We are interested in learning generative models for complex geometries described via manifolds, such as spheres, tori, and other implicit surfaces. Current extensions of existing (Euclidean) generative models are restricted to specific geometries and typically suffer from high computational costs. We introduce Moser Flow (MF), a new class of generative models within the family of continuous normalizing flows (CNF). MF also produces a CNF via a solution to the change-of-variable formula, however differently from other CNF methods, its model (learned) density is parameterized as the source (prior) density minus the divergence of a neural network (NN). The divergence is a local, linear differential operator, easy to approximate and calculate on manifolds. Therefore, unlike other CNFs, MF does not require invoking or backpropagating through an ODE solver during training. Furthermore, representing the model density explicitly as the divergence of a NN rather than as a solution of an ODE facilitates learning high fidelity densities. Theoretically, we prove that MF constitutes a universal density approximator under suitable assumptions. Empirically, we demonstrate for the first time the use of flow models for sampling from general curved surfaces and achieve significant improvements in density estimation, sample quality, and training complexity over existing CNFs on challenging synthetic geometries and real-world benchmarks from the earth and climate sciences.
****
Structure-Aware Random Fourier Kernel for Graphs
Jinyuan Fang, Qiang Zhang, Zaiqiao Meng, Shangsong Liang
Gaussian Processes (GPs) define distributions over functions and their generalization capabilities depend heavily on the choice of kernels. In this paper, we propose a novel structure-aware random Fourier (SRF) kernel for GPs that brings several benefits when modeling graph-structured data. First, SRF kernel is defined with a spectral distribution based on the Fourier duality given by the Bochner'

s theorem, transforming the kernel learning problem to a distribution inference problem. Second, SRF kernel admits a random Fourier feature formulation that makes the kernel scalable for optimization. Third, SRF kernel enables to leverage geometric structures by taking subgraphs as inputs. To effectively optimize GPs with SRF kernel, we develop a variational EM algorithm, which alternates between an inference procedure (E-step) and a learning procedure (M-step). Experimental results on five real-world datasets show that our model can achieve state-of-the-art performance in two typical graph learning tasks, i.e., object classification and link prediction.
****
Diffusion Schrödinger Bridge with Applications to Score-Based Generative Modeling
Valentin De Bortoli, James Thornton, Jeremy Heng, Arnaud Doucet
Progressively applying Gaussian noise transforms complex data distributions to approximately Gaussian. Reversing this dynamic defines a generative model. When the forward noising process is given by a Stochastic Differential Equation (SDE), Song et al (2021) demonstrate how the time inhomogeneous drift of the associated reverse-time SDE may be estimated using score-matching. A limitation of this approach is that the forward-time SDE must be run for a sufficiently long time for the final distribution to be approximately Gaussian. In contrast, solving the Schrödinger Bridge (SB) problem, i.e. an entropy-regularized optimal transport problem on path spaces, yields diffusions which generate samples from the data distribution in finite time. We present Diffusion SB (DSB), an original approximation of the Iterative Proportional Fitting (IPF) procedure to solve the SB problem, and provide theoretical analysis along with generative modeling experiments. The first DSB iteration recovers the methodology proposed by Song et al. (2021), with the flexibility of using shorter time intervals, as subsequent DSB iterations reduce the discrepancy between the final-time marginal of the forward (resp. backward) SDE with respect to the prior (resp. data) distribution. Beyond generative modeling, DSB offers a widely applicable computational optimal transport tool as the continuous state-space analogue of the popular Sinkhorn algorithm (Cuturi, 2013).
****
Improving Transferability of Representations via Augmentation-Aware Self-Supervision
Hankook Lee, Kibok Lee, Kimin Lee, Honglak Lee, Jinwoo Shin
Recent unsupervised representation learning methods have shown to be effective in a range of vision tasks by learning representations invariant to data augmentations such as random cropping and color jittering. However, such invariance could be harmful to downstream tasks if they rely on the characteristics of the data augmentations, e.g., location- or color-sensitive. This is not an issue just for unsupervised learning; we found that this occurs even in supervised learning because it also learns to predict the same label for all augmented samples of an instance. To avoid such failures and obtain more generalizable representations, we suggest to optimize an auxiliary self-supervised loss, coined AugSelf, that learns the difference of augmentation parameters (e.g., cropping positions, color adjustment intensities) between two randomly augmented samples. Our intuition is that AugSelf encourages to preserve augmentation-aware information in learned representations, which could be beneficial for their transferability. Furthermore, AugSelf can easily be incorporated into recent state-of-the-art representation learning methods with a negligible additional training cost. Extensive experiments demonstrate that our simple idea consistently improves the transferability of representations learned by supervised and unsupervised methods in various transfer learning scenarios. The code is available at https://github.com/hankook/AugSelf.
****
Long-Short Transformer: Efficient Transformers for Language and Vision
Chen Zhu, Wei Ping, Chaowei Xiao, Mohammad Shoeybi, Tom Goldstein, Anima Anandkumar, Bryan Catanzaro
Transformers have achieved success in both language and vision domains. However,

it is prohibitively expensive to scale them to long sequences such as long docu
ments or high-resolution images, because self-attention mechanism has quadratic
time and memory complexities with respect to the input sequence length. In this
paper, we propose Long-Short Transformer (Transformer-LS), an efficient self-att
ention mechanism for modeling long sequences with linear complexity for both lan
guage and vision tasks. It aggregates a novel long-range attention with dynamic
projection to model distant correlations and a short-term attention to capture f
ine-grained local correlations. We propose a dual normalization strategy to acco
unt for the scale mismatch between the two attention mechanisms. Transformer-LS
can be applied to both autoregressive and bidirectional models without additiona
l complexity. Our method outperforms the state-of-the-art models on multiple tas
ks in language and vision domains, including the Long Range Arena benchmark, aut
oregressive language modeling, and ImageNet classification. For instance, Transf
ormer-LS achieves 0.97 test BPC on enwik8 using half the number of parameters th
an previous method, while being faster and is able to handle 3x as long sequence
s compared to its full-attention version on the same hardware. On ImageNet, it c
an obtain the state-of-the-art results (e.g., a moderate size of 55.8M model sol
ely trained on 224x224 ImageNet-1K can obtain Top-1 accuracy 84.1%), while being
more scalable on high-resolution images. The source code and models are release
d at https://github.com/NVIDIA/transformer-ls.
****
Post-Training Sparsity-Aware Quantization
Gil Shomron, Freddy Gabbay, Samer Kurzum, Uri Weiser
Quantization is a technique used in deep neural networks (DNNs) to increase exec
ution performance and hardware efficiency. Uniform post-training quantization (P
TQ) methods are common, since they can be implemented efficiently in hardware an
d do not require extensive hardware resources or a training set. Mapping FP32 mo
dels to INT8 using uniform PTQ yields models with negligible accuracy degradatio
n; however, reducing precision below 8 bits with PTQ is challenging, as accuracy
degradation becomes noticeable, due to the increase in quantization noise. In t
his paper, we propose a sparsity-aware quantization (SPARQ) method, in which the
unstructured and dynamic activation sparsity is leveraged in different represen
tation granularities. 4-bit quantization, for example, is employed by dynamicall
y examining the bits of 8-bit values and choosing a window of 4 bits, while firs
t skipping zero-value bits. Moreover, instead of quantizing activation-by-activa
tion to 4 bits, we focus on pairs of 8-bit activations and examine whether one o
f the two is equal to zero. If one is equal to zero, the second can opportunisti
cally use the other's 4-bit budget; if both do not equal zero, then each is dyna
mically quantized to 4 bits, as described. SPARQ achieves minor accuracy degrada
tion and a practical hardware implementation.
****
The Implicit Bias of Minima Stability: A View from Function Space
Rotem Mulayoff, Tomer Michaeli, Daniel Soudry
Requests for name changes in the electronic proceedings will be accepted with no
questions asked. However name changes may cause bibliographic tracking issues.
Authors are asked to consider this carefully and discuss it with their co-auth
ors prior to requesting a name change in the electronic proceedings.
****
Breaking the Sample Complexity Barrier to Regret-Optimal Model-Free Reinforcemen
t Learning
Gen Li, Laixi Shi, Yuxin Chen, Yuantao Gu, Yuejie Chi
Requests for name changes in the electronic proceedings will be accepted with 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.
****
Robust Auction Design in the Auto-bidding World
Santiago Balseiro, Yuan Deng, Jieming Mao, Vahab Mirrokni, Song Zuo
In classic auction theory, reserve prices are known to be effective for improvin
g revenue for the auctioneer against quasi-linear utility maximizing bidders. Th

e introduction of reserve prices, however, usually do not help improve total wel fare of the auctioneer and the bidders. In this paper, we focus on value maximiz ing bidders with return on spend constraints---a paradigm that has drawn conside rable attention recently as more advertisers adopt auto-bidding algorithms in ad vertising platforms---and show that the introduction of reserve prices has a nov el impact on the market. Namely, by choosing reserve prices appropriately the au ctioneer can improve not only the total revenue but also the total welfare. Our results also demonstrate that reserve prices are robust to bidder types, i.e., r eserve prices work well for different bidder types, such as value maximizers and utility maximizers, without using bidder type information. We generalize these results for a variety of auction mechanisms such as VCG, GSP, and first-price au ctions. Moreover, we show how to combine these results with additive boosts to i mprove the welfare of the outcomes of the auction further. Finally, we complemen t our theoretical observations with an empirical study confirming the effectiven ess of these ideas using data from online advertising auctions.
****
Weighted model estimation for offline model-based reinforcement learning
Toru Hishinuma, Kei Senda
This paper discusses model estimation in offline model-based reinforcement learn ing (MBRL), which is important for subsequent policy improvement using an estima ted model. From the viewpoint of covariate shift, a natural idea is model estima tion weighted by the ratio of the state-action distributions of offline data and real future data. However, estimating such a natural weight is one of the main challenges for off-policy evaluation, which is not easy to use. As an artificial alternative, this paper considers weighting with the state-action distribution ratio of offline data and simulated future data, which can be estimated relative ly easily by standard density ratio estimation techniques for supervised learnin g. Based on the artificial weight, this paper defines a loss function for offlin e MBRL and presents an algorithm to optimize it. Weighting with the artificial w eight is justified as evaluating an upper bound of the policy evaluation error. Numerical experiments demonstrate the effectiveness of weighting with the artifi cial weight.
****
Practical, Provably-Correct Interactive Learning in the Realizable Setting: The Power of True Believers
Julian Katz-Samuels, Blake Mason, Kevin G. Jamieson, Rob Nowak
We consider interactive learning in the realizable setting and develop a general framework to handle problems ranging from best arm identification to active cla ssification. We begin our investigation with the observation that agnostic algor ithms \emph{cannot} be minimax-optimal in the realizable setting. Hence, we desi gn novel computationally efficient algorithms for the realizable setting that ma tch the minimax lower bound up to logarithmic factors and are general-purpose, a ccommodating a wide variety of function classes including kernel methods, H{\"o} lder smooth functions, and convex functions. The sample complexities of our algo rithms can be quantified in terms of well-known quantities like the extended tea ching dimension and haystack dimension. However, unlike algorithms based directl y on those combinatorial quantities, our algorithms are computationally efficien t. To achieve computational efficiency, our algorithms sample from the version s pace using Monte Carlo ``hit-and-run'' algorithms instead of maintaining the ver sion space explicitly. Our approach has two key strengths. First, it is simple, consisting of two unifying, greedy algorithms. Second, our algorithms have the c apability to seamlessly leverage prior knowledge that is often available and use ful in practice. In addition to our new theoretical results, we demonstrate empi rically that our algorithms are competitive with Gaussian process UCB methods.
****
Deconditional Downscaling with Gaussian Processes
Siu Lun Chau, Shahine Bouabid, Dino Sejdinovic
Refining low-resolution (LR) spatial fields with high-resolution (HR) informatio n, often known as statistical downscaling, is challenging as the diversity of sp atial datasets often prevents direct matching of observations. Yet, when LR samp

les are modeled as aggregate conditional means of HR samples with respect to a mediating variable that is globally observed, the recovery of the underlying fine-grained field can be framed as taking an "inverse" of the conditional expectation, namely a deconditioning problem. In this work, we propose a Bayesian formulation of deconditioning which naturally recovers the initial reproducing kernel Hilbert space formulation from Hsu and Ramos (2019). We extend deconditioning to a downscaling setup and devise efficient conditional mean embedding estimator for multiresolution data. By treating conditional expectations as inter-domain features of the underlying field, a posterior for the latent field can be established as a solution to the deconditioning problem. Furthermore, we show that this solution can be viewed as a two-staged vector-valued kernel ridge regressor and show that it has a minimax optimal convergence rate under mild assumptions. Lastly, we demonstrate its proficiency in a synthetic and a real-world atmospheric field downscaling problem, showing substantial improvements over existing methods.
****

Image Generation using Continuous Filter Atoms
Ze Wang, Seunghyun Hwang, Zichen Miao, Qiang Qiu
In this paper, we model the subspace of convolutional filters with a neural ordinary differential equation (ODE) to enable gradual changes in generated images. Decomposing convolutional filters over a set of filter atoms allows efficiently modeling and sampling from a subspace of high-dimensional filters. By further modeling filters atoms with a neural ODE, we show both empirically and theoretically that such introduced continuity can be propagated to the generated images, and thus achieves gradually evolved image generation. We support the proposed framework of image generation with continuous filter atoms using various experiments, including image-to-image translation and image generation conditioned on continuous labels. Without auxiliary network components and heavy supervision, the proposed continuous filter atoms allow us to easily manipulate the gradual change of generated images by controlling integration intervals of neural ordinary differential equation. This research sheds the light on using the subspace of network parameters to navigate the diverse appearance of image generation.
****

Latent Equilibrium: A unified learning theory for arbitrarily fast computation with arbitrarily slow neurons
Paul Haider, Benjamin Ellenberger, Laura Kriener, Jakob Jordan, Walter Senn, Mihai A. Petrovici
The response time of physical computational elements is finite, and neurons are no exception. In hierarchical models of cortical networks each layer thus introduces a response lag. This inherent property of physical dynamical systems results in delayed processing of stimuli and causes a timing mismatch between network output and instructive signals, thus afflicting not only inference, but also learning. We introduce Latent Equilibrium, a new framework for inference and learning in networks of slow components which avoids these issues by harnessing the ability of biological neurons to phase-advance their output with respect to their membrane potential. This principle enables quasi-instantaneous inference independent of network depth and avoids the need for phased plasticity or computationally expensive network relaxation phases. We jointly derive disentangled neuron and synapse dynamics from a prospective energy function that depends on a network's generalized position and momentum. The resulting model can be interpreted as a biologically plausible approximation of error backpropagation in deep cortical networks with continuous-time, leaky neuronal dynamics and continuously active, local plasticity. We demonstrate successful learning of standard benchmark datasets, achieving competitive performance using both fully-connected and convolutional architectures, and show how our principle can be applied to detailed models of cortical microcircuitry. Furthermore, we study the robustness of our model to spatio-temporal substrate imperfections to demonstrate its feasibility for physical realization, be it in vivo or in silico.
****

Learning Fast-Inference Bayesian Networks
Vaidyanathan Peruvemba Ramaswamy, Stefan Szeider

We propose new methods for learning Bayesian networks (BNs) that reliably support fast inference. We utilize maximum state space size as a more fine-grained measure for the BN's reasoning complexity than the standard treewidth measure, thereby accommodating the possibility that variables range over domains of different sizes. Our methods combine heuristic BN structure learning algorithms with the recently introduced MaxSAT-powered local improvement method (Peruvemba Ramaswamy and Szeider, AAAI'21). Our experiments show that our new learning methods produce BNs that support significantly faster exact probabilistic inference than BNs learned with treewidth bounds.
****

Per-Pixel Classification is Not All You Need for Semantic Segmentation
Bowen Cheng, Alex Schwing, Alexander Kirillov
Modern approaches typically formulate semantic segmentation as a per-pixel classification task, while instance-level segmentation is handled with an alternative mask classification. Our key insight: mask classification is sufficiently general to solve both semantic- and instance-level segmentation tasks in a unified manner using the exact same model, loss, and training procedure. Following this observation, we propose MaskFormer, a simple mask classification model which predicts a set of binary masks, each associated with a single global class label prediction. Overall, the proposed mask classification-based method simplifies the landscape of effective approaches to semantic and panoptic segmentation tasks and shows excellent empirical results. In particular, we observe that MaskFormer outperforms per-pixel classification baselines when the number of classes is large. Our mask classification-based method outperforms both current state-of-the-art semantic (55.6 mIoU on ADE20K) and panoptic segmentation (52.7 PQ on COCO) models.
****

Deep Markov Factor Analysis: Towards Concurrent Temporal and Spatial Analysis of fMRI Data
Amirreza Farnoosh, Sarah Ostadabbas
Factor analysis methods have been widely used in neuroimaging to transfer high dimensional imaging data into low dimensional, ideally interpretable representations. However, most of these methods overlook the highly nonlinear and complex temporal dynamics of neural processes when factorizing their imaging data. In this paper, we present deep Markov factor analysis (DMFA), a generative model that employs Markov property in a chain of low dimensional temporal embeddings together with spatial inductive assumptions, all related through neural networks, to capture temporal dynamics in functional magnetic resonance imaging (fMRI) data, and tackle their high spatial dimensionality, respectively. Augmented with a discrete latent, DMFA is able to cluster fMRI data in its low dimensional temporal embedding with regard to subject and cognitive state variability, therefore, enables validation of a variety of fMRI-driven neuroscientific hypotheses. Experimental results on both synthetic and real fMRI data demonstrate the capacity of DMFA in revealing interpretable clusters and capturing nonlinear temporal dependencies in these high dimensional imaging data.
****

BooVAE: Boosting Approach for Continual Learning of VAE
Evgenii Egorov, Anna Kuzina, Evgeny Burnaev
Variational autoencoder (VAE) is a deep generative model for unsupervised learning, allowing to encode observations into the meaningful latent space. VAE is prone to catastrophic forgetting when tasks arrive sequentially, and only the data for the current one is available. We address this problem of continual learning for VAEs. It is known that the choice of the prior distribution over the latent space is crucial for VAE in the non-continual setting. We argue that it can also be helpful to avoid catastrophic forgetting. We learn the approximation of the aggregated posterior as a prior for each task. This approximation is parametrised as an additive mixture of distributions induced by an encoder evaluated at trainable pseudo-inputs. We use a greedy boosting-like approach with entropy regularisation to learn the components. This method encourages components diversity, which is essential as we aim at memorising the current task with the fewest compo

nents possible. Based on the learnable prior, we introduce an end-to-end approac
h for continual learning of VAEs and provide empirical studies on commonly used
benchmarks (MNIST, Fashion MNIST, NotMNIST) and CelebA datasets. For each datase
t, the proposed method avoids catastrophic forgetting in a fully automatic way.
****

Handling Long-tailed Feature Distribution in AdderNets
Minjing Dong, Yunhe Wang, Xinghao Chen, Chang Xu
Adder neural networks (ANNs) are designed for low energy cost which replace expe
nsive multiplications in convolutional neural networks (CNNs) with cheaper addit
ions to yield energy-efficient neural networks and hardware accelerations. Altho
ugh ANNs achieve satisfactory efficiency, there exist gaps between ANNs and CNNs
 where the accuracy of ANNs can hardly be compared to CNNs without the assistanc
e of other training tricks, such as knowledge distillation. The inherent discrep
ancy lies in the similarity measurement between filters and features, however ho
w to alleviate this difference remains unexplored. To locate the potential probl
em of ANNs, we focus on the property difference due to similarity measurement. W
e demonstrate that unordered heavy tails in ANNs could be the key component whic
h prevents ANNs from achieving superior classification performance since fatter
tails tend to overlap in feature space. Through pre-defining Multivariate Skew L
aplace distributions and embedding feature distributions into the loss function,
 ANN features can be fully controlled and designed for various properties. We fu
rther present a novel method for tackling existing heavy tails in ANNs with only
 a modification of classifier where ANN features are clustered with their tails
well-formulated through proposed angle-based constraint on the distribution para
meters to encourage high diversity of tails. Experiments conducted on several be
nchmarks and comparison with other distributions demonstrate the effectiveness o
f proposed approach for boosting the performance of ANNs.
****

Pessimism Meets Invariance: Provably Efficient Offline Mean-Field Multi-Agent RL
Minshuo Chen, Yan Li, Ethan Wang, Zhuoran Yang, Zhaoran Wang, Tuo Zhao
Requests for name changes in the electronic proceedings will be accepted with no
 questions asked. However name changes may cause bibliographic tracking issues.
  Authors are asked to consider this carefully and discuss it with their co-auth
ors prior to requesting a name change in the electronic proceedings.
****

A Law of Iterated Logarithm for Multi-Agent Reinforcement Learning
Gugan Chandrashekhar Thoppe, Bhumesh Kumar
Requests for name changes in the electronic proceedings will be accepted with no
 questions asked. However name changes may cause bibliographic tracking issues.
  Authors are asked to consider this carefully and discuss it with their co-auth
ors prior to requesting a name change in the electronic proceedings.
****

MOMA: Multi-Object Multi-Actor Activity Parsing
Zelun Luo, Wanze Xie, Siddharth Kapoor, Yiyun Liang, Michael Cooper, Juan Carlos
 Niebles, Ehsan Adeli, Fei-Fei Li
Complex activities often involve multiple humans utilizing different objects to
complete actions (e.g., in healthcare settings, physicians, nurses, and patients
 interact with each other and various medical devices). Recognizing activities p
oses a challenge that requires a detailed understanding of actors' roles, object
s' affordances, and their associated relationships. Furthermore, these purposefu
l activities are composed of multiple achievable steps, including sub-activities
 and atomic actions, which jointly define a hierarchy of action parts. This pape
r introduces Activity Parsing as the overarching task of temporal segmentation a
nd classification of activities, sub-activities, atomic actions, along with an i
nstance-level understanding of actors, objects, and their relationships in video
s. Involving multiple entities (actors and objects), we argue that traditional p
air-wise relationships, often used in scene or action graphs, do not appropriate
ly represent the dynamics between them. Hence, we introduce Action Hypergraph, a
 spatial-temporal graph containing hyperedges (i.e., edges with higher-order rel
ationships), as a new representation. In addition, we introduce Multi-Object Mul

ti-Actor (MOMA), the first benchmark and dataset dedicated to activity parsing. Lastly, to parse a video, we propose the HyperGraph Activity Parsing (HGAP) network, which outperforms several baselines, including those based on regular graphs and raw video data.
****
The Pareto Frontier of model selection for general Contextual Bandits
Teodor Vanislavov Marinov, 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.
****
Teaching an Active Learner with Contrastive Examples
Chaoqi Wang, Adish Singla, 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.
****
Structured Denoising Diffusion Models in Discrete State-Spaces
Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, Rianne van den Berg
Denoising diffusion probabilistic models (DDPMs) [Ho et al. 2021] have shown impressive results on image and waveform generation in continuous state spaces. Here, we introduce Discrete Denoising Diffusion Probabilistic Models (D3PMs), diffusion-like generative models for discrete data that generalize the multinomial diffusion model of Hoogeboom et al. [2021], by going beyond corruption processes with uniform transition probabilities. This includes corruption with transition matrices that mimic Gaussian kernels in continuous space, matrices based on nearest neighbors in embedding space, and matrices that introduce absorbing states. The third allows us to draw a connection between diffusion models and autoregressive and mask-based generative models. We show that the choice of transition matrix is an important design decision that leads to improved results in image and text domains. We also introduce a new loss function that combines the variational lower bound with an auxiliary cross entropy loss. For text, this model class achieves strong results on character-level text generation while scaling to large vocabularies on LM1B. On the image dataset CIFAR-10, our models approach the sample quality and exceed the log-likelihood of the continuous-space DDPM model.
****
Emergent Communication of Generalizations
Jesse Mu, Noah Goodman
To build agents that can collaborate effectively with others, recent research has trained artificial agents to communicate with each other in Lewis-style referential games. However, this often leads to successful but uninterpretable communication. We argue that this is due to the game objective: communicating about a single object in a shared visual context is prone to overfitting and does not encourage language useful beyond concrete reference. In contrast, human language conveys a rich variety of abstract ideas. To promote such skills, we propose games that require communicating generalizations over sets of objects representing abstract visual concepts, optionally with separate contexts for each agent. We find that these games greatly improve systematicity and interpretability of the learned languages, according to several metrics in the literature. Finally, we propose a method for identifying logical operations embedded in the emergent languages by learning an approximate compositional reconstruction of the language.
****
Distributed Machine Learning with Sparse Heterogeneous Data
Dominic Richards, Sahand Negahban, Patrick Rebeschini
Motivated by distributed machine learning settings such as Federated Learning, we consider the problem of fitting a statistical model across a distributed collection of heterogeneous data sets whose similarity structure is encoded by a graph topology. Precisely, we analyse the case where each node is associated with fi

tting a sparse linear model, and edges join two nodes if the difference of their solutions is also sparse. We propose a method based on Basis Pursuit Denoising with a total variation penalty, and provide finite sample guarantees for sub-Gaussian design matrices. Taking the root of the tree as a reference node, we show that if the sparsity of the differences across nodes is smaller than the sparsity at the root, then recovery is successful with fewer samples than by solving the problems independently, or by using methods that rely on a large overlap in the signal supports, such as the group Lasso. We consider both the noiseless and noisy setting, and numerically investigate the performance of distributed methods based on Distributed Alternating Direction Methods of Multipliers (ADMM) and hyperspectral unmixing.
****
Manipulating SGD with Data Ordering Attacks
I Shumailov, Zakhar Shumaylov, Dmitry Kazhdan, Yiren Zhao, Nicolas Papernot, Murat A. Erdogdu, Ross J Anderson
Machine learning is vulnerable to a wide variety of attacks. It is now well understood that by changing the underlying data distribution, an adversary can poison the model trained with it or introduce backdoors. In this paper we present a novel class of training-time attacks that require no changes to the underlying dataset or model architecture, but instead only change the order in which data are supplied to the model. In particular, we find that the attacker can either prevent the model from learning, or poison it to learn behaviours specified by the attacker. Furthermore, we find that even a single adversarially-ordered epoch can be enough to slow down model learning, or even to reset all of the learning progress. Indeed, the attacks presented here are not specific to the model or dataset, but rather target the stochastic nature of modern learning procedures. We extensively evaluate our attacks on computer vision and natural language benchmarks to find that the adversary can disrupt model training and even introduce backdoors.
****
Graph Posterior Network: Bayesian Predictive Uncertainty for Node Classification
Maximilian Stadler, Bertrand Charpentier, Simon Geisler, Daniel Zügner, Stephan Günnemann
The interdependence between nodes in graphs is key to improve class prediction on nodes, utilized in approaches like Label Probagation (LP) or in Graph Neural Networks (GNNs). Nonetheless, uncertainty estimation for non-independent node-level predictions is under-explored. In this work, we explore uncertainty quantification for node classification in three ways: (1) We derive three axioms explicitly characterizing the expected predictive uncertainty behavior in homophilic attributed graphs.(2) We propose a new model Graph Posterior Network (GPN) which explicitly performs Bayesian posterior updates for predictions on interdependent nodes. GPN provably obeys the proposed axioms. (3) We extensively evaluate GPN and a strong set of baselines on semi-supervised node classification including detection of anomalous features, and detection of left-out classes. GPN outperforms existing approaches for uncertainty estimation in the experiments.
****
Locality Sensitive Teaching
Zhaozhuo Xu, Beidi Chen, Chaojian Li, Weiyang Liu, Le Song, Yingyan Lin, Anshumali Shrivastava
The emergence of the Internet-of-Things (IoT) sheds light on applying the machine teaching (MT) algorithms for online personalized education on home devices. This direction becomes more promising during the COVID-19 pandemic when in-person education becomes infeasible. However, as one of the most influential and practical MT paradigms, iterative machine teaching (IMT) is prohibited on IoT devices due to its inefficient and unscalable algorithms. IMT is a paradigm where a teacher feeds examples iteratively and intelligently based on the learner's status. In each iteration, current IMT algorithms greedily traverse the whole training set to find an example for the learner, which is computationally expensive in practice. We propose a novel teaching framework, Locality Sensitive Teaching (LST), based on locality sensitive sampling, to overcome these challenges. LST has pr

ovable near-constant time complexity, which is exponentially better than the exi sting baseline. With at most 425.12x speedups and 99.76% energy savings over IMT , LST is the first algorithm that enables energy and time efficient machine teac hing on IoT devices. Owing to LST's substantial efficiency and scalability, it i s readily applicable in real-world education scenarios.
****

No-Press Diplomacy from Scratch
Anton Bakhtin, David Wu, Adam Lerer, Noam Brown
Prior AI successes in complex games have largely focused on settings with at mos t hundreds of actions at each decision point. In contrast, Diplomacy is a game w ith more than 10^20 possible actions per turn. Previous attempts to address game s with large branching factors, such as Diplomacy, StarCraft, and Dota, used hum an data to bootstrap the policy or used handcrafted reward shaping. In this pape r, we describe an algorithm for action exploration and equilibrium approximation in games with combinatorial action spaces. This algorithm simultaneously perfor ms value iteration while learning a policy proposal network. A double oracle ste p is used to explore additional actions to add to the policy proposals. At each state, the target state value and policy for the model training are computed via an equilibrium search procedure. Using this algorithm, we train an agent, DORA, completely from scratch for a popular two-player variant of Diplomacy and show that it achieves superhuman performance. Additionally, we extend our methods to full-scale no-press Diplomacy and for the first time train an agent from scratch with no human data. We present evidence that this agent plays a strategy that i s incompatible with human-data bootstrapped agents. This presents the first stro ng evidence of multiple equilibria in Diplomacy and suggests that self play alon e may be insufficient for achieving superhuman performance in Diplomacy.
****

Remember What You Want to Forget: Algorithms for Machine Unlearning
Ayush Sekhari, Jayadev Acharya, Gautam Kamath, Ananda Theertha Suresh
Requests for name changes in the electronic proceedings will be accepted with 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 latent causal graphs via mixture oracles
Bohdan Kivva, Goutham Rajendran, Pradeep Ravikumar, Bryon Aragam
We study the problem of reconstructing a causal graphical model from data in the presence of latent variables. The main problem of interest is recovering the ca usal structure over the latent variables while allowing for general, potentially nonlinear dependencies. In many practical problems, the dependence between raw observations (e.g. pixels in an image) is much less relevant than the dependence between certain high-level, latent features (e.g. concepts or objects), and thi s is the setting of interest. We provide conditions under which both the latent representations and the underlying latent causal model are identifiable by a red uction to a mixture oracle. These results highlight an intriguing connection bet ween the well-studied problem of learning the order of a mixture model and the p roblem of learning the bipartite structure between observables and unobservables . The proof is constructive, and leads to several algorithms for explicitly reco nstructing the full graphical model. We discuss efficient algorithms and provide experiments illustrating the algorithms in practice.
****

ErrorCompensatedX: error compensation for variance reduced algorithms
Hanlin Tang, Yao Li, Ji Liu, Ming Yan
Communication cost is one major bottleneck for the scalability for distributed l earning. One approach to reduce the communication cost is to compress the gradie nt during communication. However, directly compressing the gradient decelerates the convergence speed, and the resulting algorithm may diverge for biased compre ssion. Recent work addressed this problem for stochastic gradient descent by add ing back the compression error from the previous step. This idea was further ext ended to one class of variance reduced algorithms, where the variance of the sto

chastic gradient is reduced by taking a moving average over all history gradients. However, our analysis shows that just adding the previous step's compression error, as done in existing work, does not fully compensate the compression error. So, we propose ErrorCompensateX, which uses the compression error from the previous two steps. We show that ErrorCompensateX can achieve the same asymptotic convergence rate with the training without compression. Moreover, we provide a unified theoretical analysis framework for this class of variance reduced algorithms, with or without error compensation.
****
Deep Contextual Video Compression
Jiahao Li, Bin Li, Yan Lu
Most of the existing neural video compression methods adopt the predictive coding framework, which first generates the predicted frame and then encodes its residue with the current frame. However, as for compression ratio, predictive coding is only a sub-optimal solution as it uses simple subtraction operation to remove the redundancy across frames. In this paper, we propose a deep contextual video compression framework to enable a paradigm shift from predictive coding to conditional coding. In particular, we try to answer the following questions: how to define, use, and learn condition under a deep video compression framework. To tap the potential of conditional coding, we propose using feature domain context as condition. This enables us to leverage the high dimension context to carry rich information to both the encoder and the decoder, which helps reconstruct the high-frequency contents for higher video quality. Our framework is also extensible, in which the condition can be flexibly designed. Experiments show that our method can significantly outperform the previous state-of-the-art (SOTA) deep video compression methods. When compared with x265 using veryslow preset, we can achieve 26.0% bitrate saving for 1080P standard test videos.
****
On the Frequency Bias of Generative Models
Katja Schwarz, Yiyi Liao, Andreas Geiger
The key objective of Generative Adversarial Networks (GANs) is to generate new data with the same statistics as the provided training data. However, multiple recent works show that state-of-the-art architectures yet struggle to achieve this goal. In particular, they report an elevated amount of high frequencies in the spectral statistics which makes it straightforward to distinguish real and generated images. Explanations for this phenomenon are controversial: While most works attribute the artifacts to the generator, other works point to the discriminator. We take a sober look at those explanations and provide insights on what makes proposed measures against high-frequency artifacts effective. To achieve this, we first independently assess the architectures of both the generator and discriminator and investigate if they exhibit a frequency bias that makes learning the distribution of high-frequency content particularly problematic. Based on these experiments, we make the following four observations: 1) Different upsampling operations bias the generator towards different spectral properties. 2) Checkerboard artifacts introduced by upsampling cannot explain the spectral discrepancies alone as the generator is able to compensate for these artifacts. 3) The discriminator does not struggle with detecting high frequencies per se but rather struggles with frequencies of low magnitude. 4) The downsampling operations in the discriminator can impair the quality of the training signal it provides.In light of these findings, we analyze proposed measures against high-frequency artifacts in state-of-the-art GAN training but find that none of the existing approaches can fully resolve spectral artifacts yet. Our results suggest that there is great potential in improving the discriminator and that this could be key to match the distribution of the training data more closely.
****
Learning curves of generic features maps for realistic datasets with a teacher-student model
Bruno Loureiro, Cedric Gerbelot, Hugo Cui, Sebastian Goldt, Florent Krzakala, Marc Mezard, Lenka Zdeborová
Teacher-student models provide a framework in which the typical-case performance

of high-dimensional supervised learning can be described in closed form. The as sumptions of Gaussian i.i.d. input data underlying the canonical teacher-student model may, however, be perceived as too restrictive to capture the behaviour of realistic data sets. In this paper, we introduce a Gaussian covariate generalis ation of the model where the teacher and student can act on different spaces, ge nerated with fixed, but generic feature maps. While still solvable in a closed f orm, this generalization is able to capture the learning curves for a broad rang e of realistic data sets, thus redeeming the potential of the teacher-student fr amework. Our contribution is then two-fold: first, we prove a rigorous formula f or the asymptotic training loss and generalisation error. Second, we present a n umber of situations where the learning curve of the model captures the one of a realistic data set learned with kernel regression and classification, with out-o f-the-box feature maps such as random projections or scattering transforms, or w ith pre-learned ones - such as the features learned by training multi-layer neur al networks. We discuss both the power and the limitations of the framework.
****
It Has Potential: Gradient-Driven Denoisers for Convergent Solutions to Inverse Problems
Regev Cohen, Yochai Blau, Daniel Freedman, Ehud Rivlin
In recent years there has been increasing interest in leveraging denoisers for s olving general inverse problems. Two leading frameworks are regularization-by-de noising (RED) and plug-and-play priors (PnP) which incorporate explicit likeliho od functions with priors induced by denoising algorithms. RED and PnP have show n state-of-the-art performance in diverse imaging tasks when powerful denoisersa re used, such as convolutional neural networks (CNNs). However, the study of the ir convergence remains an active line of research. Recent works derive the conv ergence of RED and PnP methods by treating CNN denoisers as approximations for m aximum a posteriori (MAP) or minimum mean square error (MMSE) estimators. Yet, state-of-the-art denoisers cannot be interpreted as either MAPor MMSE estimators , since they typically do not exhibit symmetric Jacobians. Furthermore, obtainin g stable inverse algorithms often requires controlling the Lipschitz constant of CNN denoisers during training. Precisely enforcing this constraint is impracti cal, hence, convergence cannot be completely guaranteed. In this work, we introd uce image denoisers derived as the gradients of smooth scalar-valued deep neural networks, acting as potentials. This ensures two things: (1) the proposed denoi sers display symmetric Jacobians, allowing for MAP and MMSE estimators interpret ation; (2) the denoisers may be integrated into RED and PnP schemes with backtra cking step size, removing the need for enforcing their Lipschitz constant. To sh ow the latter, we develop a simple inversion method that utilizes the proposed d enoisers. We theoretically establish its convergence to stationary points of an underlying objective function consisting of the learned potentials. We numerical ly validate our method through various imaging experiments, showing improved res ults compared to standard RED and PnP methods, and with additional provable stab ility.
****
Training Over-parameterized Models with Non-decomposable Objectives
Harikrishna Narasimhan, Aditya K. Menon
Many modern machine learning applications come with complex and nuanced design g oals such as minimizing the worst-case error, satisfying a given precision or re call target, or enforcing group-fairness constraints. Popular techniques for opt imizing such non-decomposable objectives reduce the problem into a sequence of c ost-sensitive learning tasks, each of which is then solved by re-weighting the t raining loss with example-specific costs. We point out that the standard approac h of re-weighting the loss to incorporate label costs can produce unsatisfactory results when used to train over-parameterized models. As a remedy, we propose n ew cost- sensitive losses that extend the classical idea of logit adjustment to handle more general cost matrices. Our losses are calibrated, and can be further improved with distilled labels from a teacher model. Through experiments on ben chmark image datasets, we showcase the effectiveness of our approach in training ResNet models with common robust and constrained optimization objectives.

****
Reinforcement learning for optimization of variational quantum circuit architectures
Mateusz Ostaszewski, Lea M. Trenkwalder, Wojciech Masarczyk, Eleanor Scerri, Vedran Dunjko
The study of Variational Quantum Eigensolvers (VQEs) has been in the spotlight in recent times as they may lead to real-world applications of near-term quantum devices. However, their performance depends on the structure of the used variational ansatz, which requires balancing the depth and expressivity of the corresponding circuit. At the same time, near-term restrictions limit the depth of the circuit we can expect to run. Thus, the optimization of the VQE ansatz requires maximizing the expressivity of the circuit while maintaining low depth. In recent years, various methods for VQE structure optimization have been introduced but the capacities of machine learning to aid with this problem have not yet been extensively investigated. In this work, we propose a reinforcement learning algorithm that autonomously explores the space of possible ansatzes, identifying economic circuits which still yield accurate ground energy estimates. The algorithm uses a feedback-driven curriculum learning method that autonomously adapts the complexity of the learning problem to the current performance of the learning algorithm and it incrementally improves the accuracy of the result while minimizing the circuit depth. We showcase the performance of our algorithm on the problem of estimating the ground-state energy of lithium hydride (LiH) in various configurations. In this well-known benchmark problem, we achieve chemical accuracy and state-of-the-art results in terms of circuit depth.
****
Moshpit SGD: Communication-Efficient Decentralized Training on Heterogeneous Unreliable Devices
Max Ryabinin, Eduard Gorbunov, Vsevolod Plokhotnyuk, Gennady Pekhimenko
Training deep neural networks on large datasets can often be accelerated by using multiple compute nodes. This approach, known as distributed training, can utilize hundreds of computers via specialized message-passing protocols such as Ring All-Reduce.However, running these protocols at scale requires reliable high-speed networking that is only available in dedicated clusters.In contrast, many real-world applications, such as federated learning and cloud-based distributed training, operate on unreliable devices with unstable network bandwidth.As a result, these applications are restricted to using parameter servers or gossip-based averaging protocols.In this work, we lift that restriction by proposing Moshpit All-Reduce — an iterative averaging protocol that exponentially converges to the global average.We demonstrate the efficiency of our protocol for distributed optimization with strong theoretical guarantees.The experiments show 1.3x speedup for ResNet-50 training on ImageNet compared to competitive gossip-based strategies and 1.5x speedup when training ALBERT-large on preemptible compute nodes.
****
IRM—when it works and when it doesn't: A test case of natural language inference
Yana Dranker, He He, Yonatan Belinkov
Invariant Risk Minimization (IRM) is a recently proposed framework for out-of-distribution (o.o.d) generalization. Most of the studies on IRM so far have focused on theoretical results, toy problems, and simple models. In this work, we investigate the applicability of IRM to bias mitigation-a special case of o.o.d generalization-in increasingly naturalistic settings and deep models. Using natural language inference (NLI) as a test case, we start with a setting where both the dataset and the bias are synthetic, continue with a natural dataset and synthetic bias, and end with a fully realistic setting with natural datasets and bias. Our results show that in naturalistic settings, learning complex features in place of the bias proves to be difficult, leading to a rather small improvement over empirical risk minimization. Moreover, we find that in addition to being sensitive to random seeds, the performance of IRM also depends on several critical factors, notably dataset size, bias prevalence, and bias strength, thus limiting IRM's advantage in practical scenarios. Our results highlight key challenges in applying IRM to real-world scenarios, calling for a more naturalistic characteri

zation of the problem setup for o.o.d generalization.
****
Self-Supervised Learning Disentangled Group Representation as Feature
Tan Wang, Zhongqi Yue, Jianqiang Huang, Qianru Sun, Hanwang Zhang
A good visual representation is an inference map from observations (images) to features (vectors) that faithfully reflects the hidden modularized generative factors (semantics). In this paper, we formulate the notion of "good" representation from a group-theoretic view using Higgins' definition of disentangled representation, and show that existing Self-Supervised Learning (SSL) only disentangles simple augmentation features such as rotation and colorization, thus unable to modularize the remaining semantics. To break the limitation, we propose an iterative SSL algorithm: Iterative Partition-based Invariant Risk Minimization (IP-IRM), which successfully grounds the abstract semantics and the group acting on them into concrete contrastive learning. At each iteration, IP-IRM first partitions the training samples into two subsets that correspond to an entangled group element. Then, it minimizes a subset-invariant contrastive loss, where the invariance guarantees to disentangle the group element. We prove that IP-IRM converges to a fully disentangled representation and show its effectiveness on various benchmarks. Codes are available at https://github.com/Wangt-CN/IP-IRM.
****
SalKG: Learning From Knowledge Graph Explanations for Commonsense Reasoning
Aaron Chan, Jiashu Xu, Boyuan Long, Soumya Sanyal, Tanishq Gupta, Xiang Ren
Augmenting pre-trained language models with knowledge graphs (KGs) has achieved success on various commonsense reasoning tasks. However, for a given task instance, the KG, or certain parts of the KG, may not be useful. Although KG-augmented models often use attention to focus on specific KG components, the KG is still always used, and the attention mechanism is never explicitly taught which KG components should be used. Meanwhile, saliency methods can measure how much a KG feature (e.g., graph, node, path) influences the model to make the correct prediction, thus explaining which KG features are useful. This paper explores how saliency explanations can be used to improve KG-augmented models' performance. First, we propose to create coarse (Is the KG useful?) and fine (Which nodes/paths in the KG are useful?) saliency explanations. Second, to motivate saliency-based supervision, we analyze oracle KG-augmented models which directly use saliency explanations as extra inputs for guiding their attention. Third, we propose SalKG, a framework for KG-augmented models to learn from coarse and/or fine saliency explanations. Given saliency explanations created from a task's training set, SalKG jointly trains the model to predict the explanations, then solve the task by attending to KG features highlighted by the predicted explanations. On three popular commonsense QA benchmarks (CSQA, OBQA, CODAH) and a range of KG-augmented models, we show that SalKG can yield considerable performance gains --- up to 2.76% absolute improvement on CSQA.
****
Supervising the Transfer of Reasoning Patterns in VQA
Corentin Kervadec, Christian Wolf, Grigory Antipov, Moez Baccouche, Madiha Nadri
Methods for Visual Question Anwering (VQA) are notorious for leveraging dataset biases rather than performing reasoning, hindering generalization. It has been recently shown that better reasoning patterns emerge in attention layers of a state-of-the-art VQA model when they are trained on perfect (oracle) visual inputs. This provides evidence that deep neural networks can learn to reason when training conditions are favorable enough. However, transferring this learned knowledge to deployable models is a challenge, as much of it is lost during the transfer.We propose a method for knowledge transfer based on a regularization term in our loss function, supervising the sequence of required reasoning operations.We provide a theoretical analysis based on PAC-learning, showing that such program prediction can lead to decreased sample complexity under mild hypotheses. We also demonstrate the effectiveness of this approach experimentally on the GQA dataset and show its complementarity to BERT-like self-supervised pre-training.
****
Conformal Bayesian Computation

Edwin Fong, Chris C Holmes
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
A Unified Approach to Fair Online Learning via Blackwell Approachability
Evgenii Chzhen, Christophe Giraud, Gilles Stoltz
We provide a setting and a general approach to fair online learning with stochastic sensitive and non-sensitive contexts.The setting is a repeated game between the Player and Nature, where at each stage both pick actions based on the contexts. Inspired by the notion of unawareness, we assume that the Player can only access the non-sensitive context before making a decision, while we discuss both cases of Nature accessing the sensitive contexts and Nature unaware of the sensitive contexts. Adapting Blackwell's approachability theory to handle the case of an unknown contexts' distribution, we provide a general necessary and sufficient condition for learning objectives to be compatible with some fairness constraints. This condition is instantiated on (group-wise) no-regret and (group-wise) calibration objectives, and on demographic parity as an additional constraint. When the objective is not compatible with the constraint, the provided framework permits to characterise the optimal trade-off between the two.
****
Training Neural Networks is ER-complete
Mikkel Abrahamsen, Linda Kleist, Tillmann Miltzow
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 the Under-Coverage Bias in Uncertainty Estimation
Yu Bai, Song Mei, Huan Wang, Caiming Xiong
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Decentralized Q-learning in Zero-sum Markov Games
Muhammed Sayin, Kaiqing Zhang, David Leslie, Tamer Basar, Asuman Ozdaglar
We study multi-agent reinforcement learning (MARL) in infinite-horizon discounted zero-sum Markov games. We focus on the practical but challenging setting of decentralized MARL, where agents make decisions without coordination by a centralized controller, but only based on their own payoffs and local actions executed. The agents need not observe the opponent's actions or payoffs, possibly being even oblivious to the presence of the opponent, nor be aware of the zero-sum structure of the underlying game, a setting also referred to as radically uncoupled in the literature of learning in games. In this paper, we develop a radically uncoupled Q-learning dynamics that is both rational and convergent: the learning dynamics converges to the best response to the opponent's strategy when the opponent follows an asymptotically stationary strategy; when both agents adopt the learning dynamics, they converge to the Nash equilibrium of the game. The key challenge in this decentralized setting is the non-stationarity of the environment from an agent's perspective, since both her own payoffs and the system evolution depend on the actions of other agents, and each agent adapts her policies simultaneously and independently. To address this issue, we develop a two-timescale learning dynamics where each agent updates her local Q-function and value function estimates concurrently, with the latter happening at a slower timescale.
****
Fast Certified Robust Training with Short Warmup
Zhouxing Shi, Yihan Wang, Huan Zhang, Jinfeng Yi, 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.
****
Vector-valued Distance and Gyrocalculus on the Space of Symmetric Positive Definite Matrices
Federico López, Beatrice Pozzetti, Steve Trettel, Michael Strube, Anna Wienhard
We propose the use of the vector-valued distance to compute distances and extract geometric information from the manifold of symmetric positive definite matrices (SPD), and develop gyrovector calculus, constructing analogs of vector space operations in this curved space. We implement these operations and showcase their versatility in the tasks of knowledge graph completion, item recommendation, and question answering. In experiments, the SPD models outperform their equivalents in Euclidean and hyperbolic space. The vector-valued distance allows us to visualize embeddings, showing that the models learn to disentangle representations of positive samples from negative ones.
****
Improved Transformer for High-Resolution GANs
Long Zhao, Zizhao Zhang, Ting Chen, Dimitris Metaxas, Han 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.
****
Learning High-Precision Bounding Box for Rotated Object Detection via Kullback-Leibler Divergence
Xue Yang, Xiaojiang Yang, Jirui Yang, Qi Ming, Wentao Wang, Qi Tian, Junchi Yan
Existing rotated object detectors are mostly inherited from the horizontal detection paradigm, as the latter has evolved into a well-developed area. However, these detectors are difficult to perform prominently in high-precision detection due to the limitation of current regression loss design, especially for objects with large aspect ratios. Taking the perspective that horizontal detection is a special case for rotated object detection, in this paper, we are motivated to change the design of rotation regression loss from induction paradigm to deduction methodology, in terms of the relation between rotation and horizontal detection. We show that one essential challenge is how to modulate the coupled parameters in the rotation regression loss, as such the estimated parameters can influence to each other during the dynamic joint optimization, in an adaptive and synergetic way. Specifically, we first convert the rotated bounding box into a 2-D Gaussian distribution, and then calculate the Kullback-Leibler Divergence (KLD) between the Gaussian distributions as the regression loss. By analyzing the gradient of each parameter, we show that KLD (and its derivatives) can dynamically adjust the parameter gradients according to the characteristics of the object. For instance, it will adjust the importance (gradient weight) of the angle parameter according to the aspect ratio. This mechanism can be vital for high-precision detection as a slight angle error would cause a serious accuracy drop for large aspect ratios objects. More importantly, we have proved that KLD is scale invariant. We further show that the KLD loss can be degenerated into the popular Ln-norm loss for horizontal detection. Experimental results on seven datasets using different detectors show its consistent superiority, and codes are available at https://github.com/yangxue0827/RotationDetection.
****
On Locality of Local Explanation Models
Sahra Ghalebikesabi, Lucile Ter-Minassian, Karla DiazOrdaz, Chris C Holmes
Shapley values provide model agnostic feature attributions for model outcome at a particular instance by simulating feature absence under a global population distribution. The use of a global population can lead to potentially misleading results when local model behaviour is of interest. Hence we consider the formulation of neighbourhood reference distributions that improve the local interpretability of Shapley values. By doing so, we find that the Nadaraya-Watson estimator,

a well-studied kernel regressor, can be expressed as a self-normalised importance sampling estimator. Empirically, we observe that Neighbourhood Shapley values identify meaningful sparse feature relevance attributions that provide insight into local model behaviour, complimenting conventional Shapley analysis. They also increase on-manifold explainability and robustness to the construction of adversarial classifiers.
****
FlexMatch: Boosting Semi-Supervised Learning with Curriculum Pseudo Labeling
Bowen Zhang, Yidong Wang, Wenxin Hou, HAO WU, Jindong Wang, Manabu Okumura, Takahiro Shinozaki
The recently proposed FixMatch achieved state-of-the-art results on most semi-supervised learning (SSL) benchmarks. However, like other modern SSL algorithms, FixMatch uses a pre-defined constant threshold for all classes to select unlabeled data that contribute to the training, thus failing to consider different learning status and learning difficulties of different classes. To address this issue, we propose Curriculum Pseudo Labeling (CPL), a curriculum learning approach to leverage unlabeled data according to the model's learning status. The core of CPL is to flexibly adjust thresholds for different classes at each time step to let pass informative unlabeled data and their pseudo labels. CPL does not introduce additional parameters or computations (forward or backward propagation). We apply CPL to FixMatch and call our improved algorithm FlexMatch. FlexMatch achieves state-of-the-art performance on a variety of SSL benchmarks, with especially strong performances when the labeled data are extremely limited or when the task is challenging. For example, FlexMatch achieves 13.96% and 18.96% error rate reduction over FixMatch on CIFAR-100 and STL-10 datasets respectively, when there are only 4 labels per class. CPL also significantly boosts the convergence speed, e.g., FlexMatch can use only 1/5 training time of FixMatch to achieve even better performance. Furthermore, we show that CPL can be easily adapted to other SSL algorithms and remarkably improve their performances. We open-source our code at https://github.com/TorchSSL/TorchSSL.
****
Relative Flatness and Generalization
Henning Petzka, Michael Kamp, Linara Adilova, Cristian Sminchisescu, Mario Boley
Flatness of the loss curve is conjectured to be connected to the generalization ability of machine learning models, in particular neural networks. While it has been empirically observed that flatness measures consistently correlate strongly with generalization, it is still an open theoretical problem why and under which circumstances flatness is connected to generalization, in particular in light of reparameterizations that change certain flatness measures but leave generalization unchanged. We investigate the connection between flatness and generalization by relating it to the interpolation from representative data, deriving notions of representativeness, and feature robustness. The notions allow us to rigorously connect flatness and generalization and to identify conditions under which the connection holds. Moreover, they give rise to a novel, but natural relative flatness measure that correlates strongly with generalization, simplifies to ridge regression for ordinary least squares, and solves the reparameterization issue.
****
The Image Local Autoregressive Transformer
Chenjie Cao, Yuxin Hong, Xiang Li, Chengrong Wang, Chengming Xu, Yanwei Fu, Xiangyang Xue
Recently, AutoRegressive (AR) models for the whole image generation empowered by transformers have achieved comparable or even better performance compared to Generative Adversarial Networks (GANs). Unfortunately, directly applying such AR models to edit/change local image regions, may suffer from the problems of missing global information, slow inference speed, and information leakage of local guidance. To address these limitations, we propose a novel model -- image Local Autoregressive Transformer (iLAT), to better facilitate the locally guided image synthesis. Our iLAT learns the novel local discrete representations, by the newly proposed local autoregressive (LA) transformer of the attention mask and convolu

tion mechanism. Thus iLAT can efficiently synthesize the local image regions by
key guidance information. Our iLAT is evaluated on various locally guided image
syntheses, such as pose-guided person image synthesis and face editing. Both qua
ntitative and qualitative results show the efficacy of our model.
****
Towards Multi-Grained Explainability for Graph Neural Networks
Xiang Wang, Yingxin Wu, An Zhang, Xiangnan He, Tat-Seng Chua
When a graph neural network (GNN) made a prediction, one raises question about e
xplainability: “Which fraction of the input graph is most in■uential to the mode
l’s decision?” Producing an answer requires understanding the model’s inner work
ings in general and emphasizing the insights on the decision for the instance at
 hand. Nonetheless, most of current approaches focus only on one aspect: (1) loc
al explainability, which explains each instance independently, thus hardly exhib
its the class-wise patterns; and (2) global explainability, which systematizes t
he globally important patterns, but might be trivial in the local context. This
dichotomy limits the ■exibility and effectiveness of explainers greatly. A perfo
rmant paradigm towards multi-grained explainability is until-now lacking and thu
s a focus of our work. In this work, we exploit the pre-training and ■ne-tuning
idea to develop our explainer and generate multi-grained explanations. Speci■cal
ly, the pre-training phase accounts for the contrastivity among different classe
s, so as to highlight the class-wise characteristics from a global view; afterwa
rds, the ■ne-tuning phase adapts the explanations in the local context. Experime
nts on both synthetic and real-world datasets show the superiority of our explai
ner, in terms of AUC on explaining graph classi■cation over the leading baseline
s. Our codes and datasets are available at https://github.com/Wuyxin/ReFine.
****
Behavior From the Void: Unsupervised Active Pre-Training
Hao Liu, Pieter Abbeel
We introduce a new unsupervised pre-training method for reinforcement learning c
alled APT, which stands for Active Pre-Training. APT learns behaviors and repres
entations by actively searching for novel states in reward-free environments. Th
e key novel idea is to explore the environment by maximizing a non-parametric en
tropy computed in an abstract representation space, which avoids challenging den
sity modeling and consequently allows our approach to scale much better in envir
onments that have high-dimensional observations (e.g., image observations). We e
mpirically evaluate APT by exposing task-specific reward after a long unsupervis
ed pre-training phase. In Atari games, APT achieves human-level performance on 1
2 games and obtains highly competitive performance compared to canonical fully s
upervised RL algorithms. On DMControl suite, APT beats all baselines in terms of
 asymptotic performance and data efficiency and dramatically improves performanc
e on tasks that are extremely difficult to train from scratch.
****
Autonomous Reinforcement Learning via Subgoal Curricula
Archit Sharma, Abhishek Gupta, Sergey Levine, Karol Hausman, Chelsea Finn
Reinforcement learning (RL) promises to enable autonomous acquisition of complex
 behaviors for diverse agents. However, the success of current reinforcement lea
rning algorithms is predicated on an often under-emphasised requirement -- each
trial needs to start from a fixed initial state distribution. Unfortunately, res
etting the environment to its initial state after each trial requires substantia
l amount of human supervision and extensive instrumentation of the environment w
hich defeats the goal of autonomous acquisition of complex behaviors. In this wo
rk, we propose Value-accelerated Persistent Reinforcement Learning (VaPRL), whic
h generates a curriculum of initial states such that the agent can bootstrap on
the success of easier tasks to efficiently learn harder tasks. The agent also le
arns to reach the initial states proposed by the curriculum, minimizing the reli
ance on human interventions into the learning. We observe that VaPRL reduces the
 interventions required by three orders of magnitude compared to episodic RL whi
le outperforming prior state-of-the art methods for reset-free RL both in terms
of sample efficiency and asymptotic performance on a variety of simulated roboti
cs problems.

****
Statistically and Computationally Efficient Linear Meta-representation Learning
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-auth
ors prior to requesting a name change in the electronic proceedings.
****
Decentralized Learning in Online Queuing Systems
Flore Sentenac, Etienne Boursier, 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-auth
ors prior to requesting a name change in the electronic proceedings.
****
Explainable Semantic Space by Grounding Language to Vision with Cross-Modal Cont
rastive Learning
Yizhen Zhang, Minkyu Choi, Kuan Han, Zhongming Liu
In natural language processing, most models try to learn semantic representation
s merely from texts. The learned representations encode the “distributional sema
ntics” but fail to connect to any knowledge about the physical world. In contras
t, humans learn language by grounding concepts in perception and action and the
brain encodes “grounded semantics” for cognition. Inspired by this notion and re
cent work in vision-language learning, we design a two-stream model for groundin
g language learning in vision. The model includes a VGG-based visual stream and
a Bert-based language stream. The two streams merge into a joint representationa
l space. Through cross-modal contrastive learning, the model first learns to ali
gn visual and language representations with the MS COCO dataset. The model furth
er learns to retrieve visual objects with language queries through a cross-modal
 attention module and to infer the visual relations between the retrieved object
s through a bilinear operator with the Visual Genome dataset. After training, th
e model’s language stream is a stand-alone language model capable of embedding c
oncepts in a visually grounded semantic space. This semantic space manifests pri
ncipal dimensions explainable with human intuition and neurobiological knowledge
. Word embeddings in this semantic space are predictive of human-defined norms o
f semantic features and are segregated into perceptually distinctive clusters. F
urthermore, the visually grounded language model also enables compositional lang
uage understanding based on visual knowledge and multimodal image search with qu
eries based on images, texts, or their combinations.
****
BulletTrain: Accelerating Robust Neural Network Training via Boundary Example Mi
ning
Weizhe Hua, Yichi Zhang, Chuan Guo, Zhiru Zhang, G. Edward Suh
Neural network robustness has become a central topic in machine learning in rece
nt years. Most training algorithms that improve the model's robustness to advers
arial and common corruptions also introduce a large computational overhead, requ
iring as many as ten times the number of forward and backward passes in order to
 converge. To combat this inefficiency, we propose BulletTrain, a boundary examp
le mining technique to drastically reduce the computational cost of robust train
ing. Our key observation is that only a small fraction of examples are beneficia
l for improving robustness. BulletTrain dynamically predicts these important exa
mples and optimizes robust training algorithms to focus on the important example
s. We apply our technique to several existing robust training algorithms and ach
ieve a 2.2x speed-up for TRADES and MART on CIFAR-10 and a 1.7x speed-up for Aug
Mix on CIFAR-10-C and CIFAR-100-C without any reduction in clean and robust accu
racy.
****
Neural Distance Embeddings for Biological Sequences
Gabriele Corso, Zhitao Ying, Michal Pándy, Petar Veli■kovi■, Jure Leskovec, Piet
ro Liò

The development of data-dependent heuristics and representations for biological sequences that reflect their evolutionary distance is critical for large-scale biological research. However, popular machine learning approaches, based on continuous Euclidean spaces, have struggled with the discrete combinatorial formulation of the edit distance that models evolution and the hierarchical relationship that characterises real-world datasets. We present Neural Distance Embeddings (NeuroSEED), a general framework to embed sequences in geometric vector spaces, and illustrate the effectiveness of the hyperbolic space that captures the hierarchical structure and provides an average 38% reduction in embedding RMSE against the best competing geometry. The capacity of the framework and the significance of these improvements are then demonstrated devising supervised and unsupervised NeuroSEED approaches to multiple core tasks in bioinformatics. Benchmarked with common baselines, the proposed approaches display significant accuracy and/or runtime improvements on real-world datasets. As an example for hierarchical clustering, the proposed pretrained and from-scratch methods match the quality of competing baselines with 30x and 15x runtime reduction, respectively.
****

Fitting summary statistics of neural data with a differentiable spiking network simulator

Guillaume Bellec, Shuqi Wang, Alireza Modirshanechi, Johanni Brea, Wulfram Gerstner

Fitting network models to neural activity is an important tool in neuroscience. A popular approach is to model a brain area with a probabilistic recurrent spiking network whose parameters maximize the likelihood of the recorded activity. Although this is widely used, we show that the resulting model does not produce realistic neural activity. To correct for this, we suggest to augment the log-likelihood with terms that measure the dissimilarity between simulated and recorded activity. This dissimilarity is defined via summary statistics commonly used in neuroscience and the optimization is efficient because it relies on back-propagation through the stochastically simulated spike trains. We analyze this method theoretically and show empirically that it generates more realistic activity statistics. We find that it improves upon other fitting algorithms for spiking network models like GLMs (Generalized Linear Models) which do not usually rely on back-propagation. This new fitting algorithm also enables the consideration of hidden neurons which is otherwise notoriously hard, and we show that it can be crucial when trying to infer the network connectivity from spike recordings.
****

PerSim: Data-Efficient Offline Reinforcement Learning with Heterogeneous Agents via Personalized Simulators

Anish Agarwal, Abdullah Alomar, Varkey Alumootil, Devavrat Shah, Dennis Shen, Zhi Xu, Cindy Yang

We consider offline reinforcement learning (RL) with heterogeneous agents under severe data scarcity, i.e., we only observe a single historical trajectory for every agent under an unknown, potentially sub-optimal policy. We find that the performance of state-of-the-art offline and model-based RL methods degrade significantly given such limited data availability, even for commonly perceived "solved" benchmark settings such as "MountainCar" and "CartPole". To address this challenge, we propose PerSim, a model-based offline RL approach which first learns a personalized simulator for each agent by collectively using the historical trajectories across all agents, prior to learning a policy. We do so by positing that the transition dynamics across agents can be represented as a latent function of latent factors associated with agents, states, and actions; subsequently, we theoretically establish that this function is well-approximated by a "low-rank" decomposition of separable agent, state, and action latent functions. This representation suggests a simple, regularized neural network architecture to effectively learn the transition dynamics per agent, even with scarce, offline data. We perform extensive experiments across several benchmark environments and RL methods. The consistent improvement of our approach, measured in terms of both state dynamics prediction and eventual reward, confirms the efficacy of our framework in leveraging limited historical data to simultaneously learn personalized polici

es across agents.
****

Online Sign Identification: Minimization of the Number of Errors in Thresholding Bandits

Reda Ouhamma, Odalric-Ambrym Maillard, Vianney Perchet

In the fixed budget thresholding bandit problem, an algorithm sequentially allocates a budgeted number of samples to different distributions. It then predicts whether the mean of each distribution is larger or lower than a given threshold. We introduce a large family of algorithms (containing most existing relevant ones), inspired by the Frank-Wolfe algorithm, and provide a thorough yet generic analysis of their performance. This allowed us to construct new explicit algorithms, for a broad class of problems, whose losses are within a small constant factor of the non-adaptive oracle ones. Quite interestingly, we observed that adaptive methodsempirically greatly out-perform non-adaptive oracles, an uncommon behavior in standard online learning settings, such as regret minimization. We explain this surprising phenomenon on an insightful toy problem.
****

All Tokens Matter: Token Labeling for Training Better Vision Transformers

Zi-Hang Jiang, Qibin Hou, Li Yuan, Daquan Zhou, Yujun Shi, Xiaojie Jin, Anran Wang, Jiashi Feng

In this paper, we present token labeling---a new training objective for training high-performance vision transformers (ViTs). Different from the standard training objective of ViTs that computes the classification loss on an additional trainable class token, our proposed one takes advantage of all the image patch tokens to compute the training loss in a dense manner. Specifically, token labeling reformulates the image classification problem into multiple token-level recognition problems and assigns each patch token with an individual location-specific supervision generated by a machine annotator. Experiments show that token labeling can clearly and consistently improve the performance of various ViT models across a wide spectrum. For a vision transformer with 26M learnable parameters serving as an example, with token labeling, the model can achieve 84.4% Top-1 accuracy on ImageNet. The result can be further increased to 86.4% by slightly scaling the model size up to 150M, delivering the minimal-sized model among previous models (250M+) reaching 86%. We also show that token labeling can clearly improve the generalization capability of the pretrained models on downstream tasks with dense prediction, such as semantic segmentation. Our code and model are publicly available at https://github.com/zihangJiang/TokenLabeling.
****

Partition and Code: learning how to compress graphs

Giorgos Bouritsas, Andreas Loukas, Nikolaos Karalias, Michael Bronstein

Can we use machine learning to compress graph data? The absence of ordering in graphs poses a significant challenge to conventional compression algorithms, limiting their attainable gains as well as their ability to discover relevant patterns. On the other hand, most graph compression approaches rely on domain-dependent handcrafted representations and cannot adapt to different underlying graph distributions. This work aims to establish the necessary principles a lossless graph compression method should follow to approach the entropy storage lower bound. Instead of making rigid assumptions about the graph distribution, we formulate the compressor as a probabilistic model that can be learned from data and generalise to unseen instances. Our “Partition and Code” framework entails three steps: first, a partitioning algorithm decomposes the graph into subgraphs, then these are mapped to the elements of a small dictionary on which we learn a probability distribution, and finally, an entropy encoder translates the representation into bits. All the components (partitioning, dictionary and distribution) are parametric and can be trained with gradient descent. We theoretically compare the compression quality of several graph encodings and prove, under mild conditions, that PnC achieves compression gains that grow either linearly or quadratically with the number of vertices. Empirically, PnC yields significant compression improvements on diverse real-world networks.
****

Knowledge-inspired 3D Scene Graph Prediction in Point Cloud
Shoulong Zhang, shuai li, Aimin Hao, Hong Qin
Prior knowledge integration helps identify semantic entities and their relations hips in a graphical representation, however, its meaningful abstraction and inte rvention remain elusive. This paper advocates a knowledge-inspired 3D scene grap h prediction method solely based on point clouds. At the mathematical modeling l evel, we formulate the task as two sub-problems: knowledge learning and scene gr aph prediction with learned prior knowledge. Unlike conventional methods that le arn knowledge embedding and regular patterns from encoded visual information, we propose to suppress the misunderstandings caused by appearance similarities and other perceptual confusion. At the network design level, we devise a graph auto -encoder to automatically extract class-dependent representations and topologica l patterns from the one-hot class labels and their intrinsic graphical structure s, so that the prior knowledge can avoid perceptual errors and noises. We furthe r devise a scene graph prediction model to predict credible relationship triplet s by incorporating the related prototype knowledge with perceptual information. Comprehensive experiments confirm that, our method can successfully learn repres entative knowledge embedding, and the obtained prior knowledge can effectively e nhance the accuracy of relationship predictions. Our thorough evaluations indica te the new method can achieve the state-of-the-art performance compared with oth er scene graph prediction methods.
****
Online Variational Filtering and Parameter Learning
Andrew Campbell, Yuyang Shi, Thomas Rainforth, Arnaud Doucet
We present a variational method for online state estimation and parameter learni ng in state-space models (SSMs), a ubiquitous class of latent variable models fo r sequential data. As per standard batch variational techniques, we use stochast ic gradients to simultaneously optimize a lower bound on the log evidence with r espect to both model parameters and a variational approximation of the states' p osterior distribution. However, unlike existing approaches, our method is able t o operate in an entirely online manner, such that historic observations do not r equire revisitation after being incorporated and the cost of updates at each tim e step remains constant, despite the growing dimensionality of the joint posteri or distribution of the states. This is achieved by utilizing backward decomposit ions of this joint posterior distribution and of its variational approximation, combined with Bellman-type recursions for the evidence lower bound and its gradi ents. We demonstrate the performance of this methodology across several examples , including high-dimensional SSMs and sequential Variational Auto-Encoders.
****
Heavy Ball Neural Ordinary Differential Equations
Hedi Xia, Vai Suliafu, Hangjie Ji, Tan Nguyen, Andrea Bertozzi, Stanley Osher, B ao Wang
We propose heavy ball neural ordinary differential equations (HBNODEs), leveragi ng the continuous limit of the classical momentum accelerated gradient descent, to improve neural ODEs (NODEs) training and inference. HBNODEs have two properti es that imply practical advantages over NODEs: (i) The adjoint state of an HBNOD E also satisfies an HBNODE, accelerating both forward and backward ODE solvers, thus significantly reducing the number of function evaluations (NFEs) and improv ing the utility of the trained models. (ii) The spectrum of HBNODEs is well stru ctured, enabling effective learning of long-term dependencies from complex seque ntial data. We verify the advantages of HBNODEs over NODEs on benchmark tasks, i ncluding image classification, learning complex dynamics, and sequential modelin g. Our method requires remarkably fewer forward and backward NFEs, is more accur ate, and learns long-term dependencies more effectively than the other ODE-based neural network models. Code is available at \url{https://github.com/hedixia/Hea vyBallNODE}.
****
Structure learning in polynomial time: Greedy algorithms, Bregman information, a nd exponential families
Goutham Rajendran, Bohdan Kivva, Ming Gao, Bryon Aragam

Greedy algorithms have long been a workhorse for learning graphical models, and more broadly for learning statistical models with sparse structure. In the conte xt of learning directed acyclic graphs, greedy algorithms are popular despite th eir worst-case exponential runtime. In practice, however, they are very efficien t. We provide new insight into this phenomenon by studying a general greedy scor e-based algorithm for learning DAGs. Unlike edge-greedy algorithms such as the p opular GES and hill-climbing algorithms, our approach is vertex-greedy and requi res at most a polynomial number of score evaluations. We then show how recent po lynomial-time algorithms for learning DAG models are a special case of this algo rithm, thereby illustrating how these order-based algorithms can be rigourously interpreted as score-based algorithms. This observation suggests new score funct ions and optimality conditions based on the duality between Bregman divergences and exponential families, which we explore in detail. Explicit sample and comput ational complexity bounds are derived. Finally, we provide extensive experiments suggesting that this algorithm indeed optimizes the score in a variety of setti ngs.
****
On the Sample Complexity of Learning under Geometric Stability
Alberto Bietti, Luca Venturi, Joan Bruna
Many supervised learning problems involve high-dimensional data such as images, text, or graphs. In order to make efficient use of data, it is often useful to l everage certain geometric priors in the problem at hand, such as invariance to t ranslations, permutation subgroups, or stability to small deformations. We study the sample complexity of learning problems where the target function presents s uch invariance and stability properties, by considering spherical harmonic decom positions of such functions on the sphere. We provide non-parametric rates of co nvergence for kernel methods, and show improvements in sample complexity by a fa ctor equal to the size of the group when using an invariant kernel over the grou p, compared to the corresponding non-invariant kernel. These improvements are va lid when the sample size is large enough, with an asymptotic behavior that depen ds on spectral properties of the group. Finally, these gains are extended beyond invariance groups to also cover geometric stability to small deformations, mode led here as subsets (not necessarily subgroups) of permutations.
****
SIMILAR: Submodular Information Measures Based Active Learning In Realistic Scen arios
Suraj Kothawade, Nathan Beck, Krishnateja Killamsetty, Rishabh Iyer
Active learning has proven to be useful for minimizing labeling costs by selecting the most informative samples. However, existing active lear ning methods do not work well in realistic scenarios such as imbalance or rare c lasses,out-of-distribution data in the unlabeled set, and redundancy. In this w ork, we propose SIMILAR (Submodular Information Measures based actIve LeARning), a unified active learning framework using recently proposed submodular informat ion measures (SIM) as acquisition functions. We argue that SIMILAR not only work s in standard active learning but also easily extends to the realistic settings considered above and acts as a one-stop solution for active learning that is sca lable to large real-world datasets. Empirically, we show that SIMILAR significan tly outperforms existing active learning algorithms by as much as ~5%–18%in the case of rare classes and ~5%–10%in the case of out-of-distribution data on sever al image classification tasks like CIFAR-10, MNIST, and ImageNet.
****
Monte Carlo Tree Search With Iteratively Refining State Abstractions
Samuel Sokota, Caleb Y Ho, Zaheen Ahmad, J. Zico Kolter
Decision-time planning is the process of constructing a transient, local policy with the intent of using it to make the immediate decision. Monte Carlo tree sea rch (MCTS), which has been leveraged to great success in Go, chess, shogi, Hex, Atari, and other settings, is perhaps the most celebrated decision-time planning algorithm. Unfortunately, in its original form, MCTS can degenerate to one-step search in domains with stochasticity. Progressive widening is one way to amelio rate this issue, but we argue that it possesses undesirable properties for some

settings. In this work, we present a method, called abstraction refining, for extending MCTS to stochastic environments which, unlike progressive widening, leverages the geometry of the state space. We argue that leveraging the geometry of the space can offer advantages. To support this claim, we present a series of experimental examples in which abstraction refining outperforms progressive widening, given equal simulation budgets.

****

Flattening Sharpness for Dynamic Gradient Projection Memory Benefits Continual Learning

Danruo DENG, Guangyong Chen, Jianye Hao, Qiong Wang, Pheng-Ann Heng

The backpropagation networks are notably susceptible to catastrophic forgetting, where networks tend to forget previously learned skills upon learning new ones. To address such the 'sensitivity-stability' dilemma, most previous efforts have been contributed to minimizing the empirical risk with different parameter regularization terms and episodic memory, but rarely exploring the usages of the weight loss landscape. In this paper, we investigate the relationship between the weight loss landscape and sensitivity-stability in the continual learning scenario, based on which, we propose a novel method, Flattening Sharpness for Dynamic Gradient Projection Memory (FS-DGPM). In particular, we introduce a soft weight to represent the importance of each basis representing past tasks in GPM, which can be adaptively learned during the learning process, so that less important bases can be dynamically released to improve the sensitivity of new skill learning. We further introduce Flattening Sharpness (FS) to reduce the generalization gap by explicitly regulating the flatness of the weight loss landscape of all seen tasks. As demonstrated empirically, our proposed method consistently outperforms baselines with the superior ability to learn new skills while alleviating forgetting effectively.

****

Taxonomizing local versus global structure in neural network loss landscapes

Yaoqing Yang, Liam Hodgkinson, Ryan Theisen, Joe Zou, Joseph E. Gonzalez, Kannan Ramchandran, Michael W. Mahoney

Viewing neural network models in terms of their loss landscapes has a long history in the statistical mechanics approach to learning, and in recent years it has received attention within machine learning proper. Among other things, local metrics (such as the smoothness of the loss landscape) have been shown to correlate with global properties of the model (such as good generalization performance). Here, we perform a detailed empirical analysis of the loss landscape structure of thousands of neural network models, systematically varying learning tasks, model architectures, and/or quantity/quality of data. By considering a range of metrics that attempt to capture different aspects of the loss landscape, we demonstrate that the best test accuracy is obtained when: the loss landscape is globally well-connected; ensembles of trained models are more similar to each other; and models converge to locally smooth regions. We also show that globally poorly-connected landscapes can arise when models are small or when they are trained to lower quality data; and that, if the loss landscape is globally poorly-connected, then training to zero loss can actually lead to worse test accuracy. Our detailed empirical results shed light on phases of learning (and consequent double descent behavior), fundamental versus incidental determinants of good generalization, the role of load-like and temperature-like parameters in the learning process, different influences on the loss landscape from model and data, and the relationships between local and global metrics, all topics of recent interest.

****

Learning Models for Actionable Recourse

Alexis Ross, Himabindu Lakkaraju, Osbert Bastani

As machine learning models are increasingly deployed in high-stakes domains such as legal and financial decision-making, there has been growing interest in post-hoc methods for generating counterfactual explanations. Such explanations provide individuals adversely impacted by predicted outcomes (e.g., an applicant denied a loan) with recourse---i.e., a description of how they can change their features to obtain a positive outcome. We propose a novel algorithm that leverages a

dversarial training and PAC confidence sets to learn models that theoretically g uarantee recourse to affected individuals with high probability without sacrific ing accuracy. We demonstrate the efficacy of our approach via extensive experime nts on real data.
****
Efficient and Accurate Gradients for Neural SDEs
Patrick Kidger, James Foster, Xuechen (Chen) Li, Terry Lyons
Requests for name changes in the electronic proceedings will be accepted with 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.
****
EIGNN: Efficient Infinite-Depth Graph Neural Networks
Juncheng Liu, Kenji Kawaguchi, Bryan Hooi, Yiwei Wang, Xiaokui Xiao
Graph neural networks (GNNs) are widely used for modelling graph-structured data in numerous applications. However, with their inherently finite aggregation lay ers, existing GNN models may not be able to effectively capture long-range depen dencies in the underlying graphs. Motivated by this limitation, we propose a GNN model with infinite depth, which we call Efficient Infinite-Depth Graph Neural Networks (EIGNN), to efficiently capture very long-range dependencies. We theore tically derive a closed-form solution of EIGNN which makes training an infinite-depth GNN model tractable. We then further show that we can achieve more efficie nt computation for training EIGNN by using eigendecomposition. The empirical res ults of comprehensive experiments on synthetic and real-world datasets show that EIGNN has a better ability to capture long-range dependencies than recent basel ines, and consistently achieves state-of-the-art performance. Furthermore, we sh ow that our model is also more robust against both noise and adversarial perturb ations on node features.
****
Fractal Structure and Generalization Properties of Stochastic Optimization Algor ithms
Alexander Camuto, George Deligiannidis, Murat A. Erdogdu, Mert Gurbuzbalaban, Um ut Simsekli, Lingjiong Zhu
Understanding generalization in deep learning has been one of the major challeng es in statistical learning theory over the last decade. While recent work has il lustrated that the dataset and the training algorithm must be taken into account in order to obtain meaningful generalization bounds, it is still theoretically not clear which properties of the data and the algorithm determine the generaliz ation performance. In this study, we approach this problem from a dynamical syst ems theory perspective and represent stochastic optimization algorithms as \emph {random iterated function systems} (IFS). Well studied in the dynamical systems literature, under mild assumptions, such IFSs can be shown to be ergodic with an invariant measure that is often supported on sets with a \emph{fractal structur e}. As our main contribution, we prove that the generalization error of a stocha stic optimization algorithm can be bounded based on the `complexity' of the frac tal structure that underlies its invariant measure. Then, by leveraging results from dynamical systems theory, we show that the generalization error can be expl icitly linked to the choice of the algorithm (e.g., stochastic gradient descent -- SGD), algorithm hyperparameters (e.g., step-size, batch-size), and the geomet ry of the problem (e.g., Hessian of the loss). We further specialize our results to specific problems (e.g., linear/logistic regression, one hidden-layered neur al networks) and algorithms (e.g., SGD and preconditioned variants), and obtain analytical estimates for our bound. For modern neural networks, we develop an ef ficient algorithm to compute the developed bound and support our theory with var ious experiments on neural networks.
****
An Infinite-Feature Extension for Bayesian ReLU Nets That Fixes Their Asymptotic Overconfidence
Agustinus Kristiadi, Matthias Hein, Philipp Hennig
A Bayesian treatment can mitigate overconfidence in ReLU nets around the trainin

g data. But far away from them, ReLU Bayesian neural networks (BNNs) can still u nderestimate uncertainty and thus be asymptotically overconfident. This issue ar ises since the output variance of a BNN with finitely many features is quadratic in the distance from the data region. Meanwhile, Bayesian linear models with Re LU features converge, in the infinite-width limit, to a particular Gaussian proc ess (GP) with a variance that grows cubically so that no asymptotic overconfiden ce can occur. While this may seem of mostly theoretical interest, in this work, we show that it can be used in practice to the benefit of BNNs. We extend finite ReLU BNNs with infinite ReLU features via the GP and show that the resulting mo del is asymptotically maximally uncertain far away from the data while the BNNs' predictive power is unaffected near the data. Although the resulting model appr oximates a full GP posterior, thanks to its structure, it can be applied post-ho c to any pre-trained ReLU BNN at a low cost.
****
Bandit Phase Retrieval
Tor Lattimore, Botao Hao
Requests for name changes in the electronic proceedings will be accepted with 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.
****
Lower Bounds on Metropolized Sampling Methods for Well-Conditioned Distributions
Yin Tat Lee, Ruoqi Shen, 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-auth ors prior to requesting a name change in the electronic proceedings.
****
Taming Communication and Sample Complexities in Decentralized Policy Evaluation for Cooperative Multi-Agent Reinforcement Learning
Xin Zhang, Zhuqing Liu, Jia Liu, Zhengyuan Zhu, Songtao 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-auth ors prior to requesting a name change in the electronic proceedings.
****
Federated Graph Classification over Non-IID Graphs
Han Xie, Jing Ma, Li Xiong, Carl Yang
Federated learning has emerged as an important paradigm for training machine lea rning models in different domains. For graph-level tasks such as graph classific ation, graphs can also be regarded as a special type of data samples, which can be collected and stored in separate local systems. Similar to other domains, mul tiple local systems, each holding a small set of graphs, may benefit from collab oratively training a powerful graph mining model, such as the popular graph neur al networks (GNNs). To provide more motivation towards such endeavors, we analyz e real-world graphs from different domains to confirm that they indeed share cer tain graph properties that are statistically significant compared with random gr aphs. However, we also find that different sets of graphs, even from the same do main or same dataset, are non-IID regarding both graph structures and node featu res. To handle this, we propose a graph clustered federated learning (GCFL) fram ework that dynamically finds clusters of local systems based on the gradients of GNNs, and theoretically justify that such clusters can reduce the structure and feature heterogeneity among graphs owned by the local systems. Moreover, we obs erve the gradients of GNNs to be rather fluctuating in GCFL which impedes high-q uality clustering, and design a gradient sequence-based clustering mechanism bas ed on dynamic time warping (GCFL+). Extensive experimental results and in-depth analysis demonstrate the effectiveness of our proposed frameworks.
****
SubTab: Subsetting Features of Tabular Data for Self-Supervised Representation L earning

Talip Ucar, Ehsan Hajiramezanali, Lindsay Edwards
Self-supervised learning has been shown to be very effective in learning useful representations, and yet much of the success is achieved in data types such as images, audio, and text. The success is mainly enabled by taking advantage of spatial, temporal, or semantic structure in the data through augmentation. However, such structure may not exist in tabular datasets commonly used in fields such as healthcare, making it difficult to design an effective augmentation method, and hindering a similar progress in tabular data setting. In this paper, we introduce a new framework, Subsetting features of Tabular data (SubTab), that turns the task of learning from tabular data into a multi-view representation learning problem by dividing the input features to multiple subsets. We argue that reconstructing the data from the subset of its features rather than its corrupted version in an autoencoder setting can better capture its underlying latent representation. In this framework, the joint representation can be expressed as the aggregate of latent variables of the subsets at test time, which we refer to as collaborative inference. Our experiments show that the SubTab achieves the state of the art (SOTA) performance of 98.31% on MNIST in tabular setting, on par with CNN-based SOTA models, and surpasses existing baselines on three other real-world datasets by a significant margin.
****

Convergence Rates of Stochastic Gradient Descent under Infinite Noise Variance
Hongjian Wang, Mert Gurbuzbalaban, Lingjiong Zhu, Umut Simsekli, Murat A. Erdogdu
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

Conflict-Averse Gradient Descent for Multi-task learning
Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, Qiang Liu
The goal of multi-task learning is to enable more efficient learning than single task learning by sharing model structures for a diverse set of tasks. A standard multi-task learning objective is to minimize the average loss across all tasks. While straightforward, using this objective often results in much worse final performance for each task than learning them independently. A major challenge in optimizing a multi-task model is the conflicting gradients, where gradients of different task objectives are not well aligned so that following the average gradient direction can be detrimental to specific tasks' performance. Previous work has proposed several heuristics to manipulate the task gradients for mitigating this problem. But most of them lack convergence guarantee and/or could converge to any Pareto-stationary point.In this paper, we introduce Conflict-Averse Gradient descent (CAGrad) which minimizes the average loss function, while leveraging the worst local improvement of individual tasks to regularize the algorithm trajectory. CAGrad balances the objectives automatically and still provably converges to a minimum over the average loss. It includes the regular gradient descent (GD) and the multiple gradient descent algorithm (MGDA) in the multi-objective optimization (MOO) literature as special cases. On a series of challenging multi-task supervised learning and reinforcement learning tasks, CAGrad achieves improved performance over prior state-of-the-art multi-objective gradient manipulation methods.
****

Amortized Synthesis of Constrained Configurations Using a Differentiable Surrogate
Xingyuan Sun, Tianju Xue, Szymon Rusinkiewicz, Ryan P. Adams
In design, fabrication, and control problems, we are often faced with the task of synthesis, in which we must generate an object or configuration that satisfies a set of constraints while maximizing one or more objective functions. The synthesis problem is typically characterized by a physical process in which many different realizations may achieve the goal. This many-to-one map presents challenges to the supervised learning of feed-forward synthesis, as the set of viable de

signs may have a complex structure. In addition, the non-differentiable nature of many physical simulations prevents efficient direct optimization. We address both of these problems with a two-stage neural network architecture that we may consider to be an autoencoder. We first learn the decoder: a differentiable surrogate that approximates the many-to-one physical realization process. We then learn the encoder, which maps from goal to design, while using the fixed decoder to evaluate the quality of the realization. We evaluate the approach on two case studies: extruder path planning in additive manufacturing and constrained soft robot inverse kinematics. We compare our approach to direct optimization of the design using the learned surrogate, and to supervised learning of the synthesis problem. We find that our approach produces higher quality solutions than supervised learning, while being competitive in quality with direct optimization, at a greatly reduced computational cost.
****
Efficient First-Order Contextual Bandits: Prediction, Allocation, and Triangular Discrimination
Dylan J. Foster, Akshay Krishnamurthy
A recurring theme in statistical learning, online learning, and beyond is that faster convergence rates are possible for problems with low noise, often quantified by the performance of the best hypothesis; such results are known as first-order or small-loss guarantees. While first-order guarantees are relatively well understood in statistical and online learning, adapting to low noise in contextual bandits (and more broadly, decision making) presents major algorithmic challenges. In a COLT 2017 open problem, Agarwal, Krishnamurthy, Langford, Luo, and Schapire asked whether first-order guarantees are even possible for contextual bandits and---if so---whether they can be attained by efficient algorithms. We give a resolution to this question by providing an optimal and efficient reduction from contextual bandits to online regression with the logarithmic (or, cross-entropy) loss. Our algorithm is simple and practical, readily accommodates rich function classes, and requires no distributional assumptions beyond realizability. In a large-scale empirical evaluation, we find that our approach typically outperforms comparable non-first-order methods.On the technical side, we show that the logarithmic loss and an information-theoretic quantity called the triangular discrimination play a fundamental role in obtaining first-order guarantees, and we combine this observation with new refinements to the regression oracle reduction framework of Foster and Rakhlin (2020). The use of triangular discrimination yields novel results even for the classical statistical learning model, and we anticipate that it will find broader use.
****
Distributed Estimation with Multiple Samples per User: Sharp Rates and Phase Transition
Jayadev Acharya, Clement Canonne, Yuhan Liu, Ziteng Sun, Himanshu Tyagi
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Revisiting Deep Learning Models for Tabular Data
Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, Artem Babenko
The existing literature on deep learning for tabular data proposes a wide range of novel architectures and reports competitive results on various datasets. However, the proposed models are usually not properly compared to each other and existing works often use different benchmarks and experiment protocols. As a result, it is unclear for both researchers and practitioners what models perform best. Additionally, the field still lacks effective baselines, that is, the easy-to-use models that provide competitive performance across different problems.In this work, we perform an overview of the main families of DL architectures for tabular data and raise the bar of baselines in tabular DL by identifying two simple and powerful deep architectures. The first one is a ResNet-like architecture which turns out to be a strong baseline that is often missing in prior works. The se

cond model is our simple adaptation of the Transformer architecture for tabular data, which outperforms other solutions on most tasks. Both models are compared to many existing architectures on a diverse set of tasks under the same training and tuning protocols. We also compare the best DL models with Gradient Boosted Decision Trees and conclude that there is still no universally superior solution. The source code is available at https://github.com/yandex-research/rtdl.
****

Backdoor Attack with Imperceptible Input and Latent Modification
Khoa Doan, Yingjie Lao, Ping Li
Recent studies have shown that deep neural networks (DNN) are vulnerable to various adversarial attacks. In particular, an adversary can inject a stealthy backdoor into a model such that the compromised model will behave normally without the presence of the trigger. Techniques for generating backdoor images that are visually imperceptible from clean images have also been developed recently, which further enhance the stealthiness of the backdoor attacks from the input space. Along with the development of attacks, defense against backdoor attacks is also evolving. Many existing countermeasures found that backdoor tends to leave tangible footprints in the latent or feature space, which can be utilized to mitigate backdoor attacks.In this paper, we extend the concept of imperceptible backdoor from the input space to the latent representation, which significantly improves the effectiveness against the existing defense mechanisms, especially those relying on the distinguishability between clean inputs and backdoor inputs in latent space. In the proposed framework, the trigger function will learn to manipulate the input by injecting imperceptible input noise while matching the latent representations of the clean and manipulated inputs via a Wasserstein-based regularization of the corresponding empirical distributions. We formulate such an objective as a non-convex and constrained optimization problem and solve the problem with an efficient stochastic alternating optimization procedure. We name the proposed backdoor attack as Wasserstein Backdoor (WB), which achieves a high attack success rate while being stealthy from both the input and latent spaces, as tested in several benchmark datasets, including MNIST, CIFAR10, GTSRB, and TinyImagenet.
****

SOPE: Spectrum of Off-Policy Estimators
Christina Yuan, Yash Chandak, Stephen Giguere, Philip S. Thomas, Scott Niekum
Many sequential decision making problems are high-stakes and require off-policy evaluation (OPE) of a new policy using historical data collected using some other policy. One of the most common OPE techniques that provides unbiased estimates is trajectory based importance sampling (IS). However, due to the high variance of trajectory IS estimates, importance sampling methods based on state-action visitation distributions (SIS) have recently been adopted. Unfortunately, while SIS often provides lower variance estimates for long horizons, estimating the state-action distribution ratios can be challenging and lead to biased estimates. In this paper, we present a new perspective on this bias-variance trade-off and show the existence of a spectrum of estimators whose endpoints are SIS and IS. Additionally, we also establish a spectrum for doubly-robust and weighted version of these estimators. We provide empirical evidence that estimators in this spectrum can be used to trade-off between the bias and variance of IS and SIS and can achieve lower mean-squared error than both IS and SIS.
****

Label-Imbalanced and Group-Sensitive Classification under Overparameterization
Ganesh Ramachandra Kini, Orestis Paraskevas, Samet Oymak, Christos Thrampoulidis
The goal in label-imbalanced and group-sensitive classification is to optimize relevant metrics such as balanced error and equal opportunity. Classical methods, such as weighted cross-entropy, fail when training deep nets to the terminal phase of training (TPT), that is training beyond zero training error. This observation has motivated recent flurry of activity in developing heuristic alternatives following the intuitive mechanism of promoting larger margin for minorities. In contrast to previous heuristics, we follow a principled analysis explaining how different loss adjustments affect margins. First, we prove that for all linear

classifiers trained in TPT, it is necessary to introduce multiplicative, rather than additive, logit adjustments so that the interclass margins change appropriately. To show this, we discover a connection of the multiplicative CE modification to the cost-sensitive support-vector machines. Perhaps counterintuitively, we also find that, at the start of training, the same multiplicative weights can actually harm the minority classes. Thus, while additive adjustments are ineffective in the TPT, we show that they can speed up convergence by countering the initial negative effect of the multiplicative weights. Motivated by these findings, we formulate the vector-scaling (VS) loss, that captures existing techniques as special cases. Moreover, we introduce a natural extension of the VS-loss to group-sensitive classification, thus treating the two common types of imbalances (label/group) in a unifying way. Importantly, our experiments on state-of-the-art datasets are fully consistent with our theoretical insights and confirm the superior performance of our algorithms. Finally, for imbalanced Gaussian-mixtures data, we perform a generalization analysis, revealing tradeoffs between balanced / standard error and equal opportunity.
****

Neural Program Generation Modulo Static Analysis

Rohan Mukherjee, Yeming Wen, Dipak Chaudhari, Thomas Reps, Swarat Chaudhuri, Christopher Jermaine

State-of-the-art neural models of source code tend to be evaluated on the generation of individual expressions and lines of code, and commonly fail on long-horizon tasks such as the generation of entire method bodies. We propose to address this deficiency using weak supervision from a static program analyzer. Our neurosymbolic method allows a deep generative model to symbolically compute, using calls to a static analysis tool, long-distance semantic relationships in the code that it has already generated. During training, the model observes these relationships and learns to generate programs conditioned on them. We apply our approach to the problem of generating entire Java methods given the remainder of the class that contains the method. Our experiments show that the approach substantially outperforms a state-of-the-art transformer and a model that explicitly tries to learn program semantics on this task, both in terms of producing programs free of basic semantic errors and in terms of syntactically matching the ground truth.
****

Unfolding Taylor's Approximations for Image Restoration

man zhou, Xueyang Fu, Zeyu Xiao, Gang Yang, Aiping Liu, Zhiwei Xiong

Deep learning provides a new avenue for image restoration, which demands a delicate balance between fine-grained details and high-level contextualized information during recovering the latent clear image. In practice, however, existing methods empirically construct encapsulated end-to-end mapping networks without deepening into the rationality, and neglect the intrinsic prior knowledge of restoration task. To solve the above problems, inspired by Taylor’s Approximations, we unfold Taylor’s Formula to construct a novel framework for image restoration. We find the main part and the derivative part of Taylor’s Approximations take the same effect as the two competing goals of high-level contextualized information and spatial details of image restoration respectively. Specifically, our framework consists of two steps, which are correspondingly responsible for the mapping and derivative functions. The former first learns the high-level contextualized information and the later combines it with the degraded input to progressively recover local high-order spatial details. Our proposed framework is orthogonal to existing methods and thus can be easily integrated with them for further improvement, and extensive experiments demonstrate the effectiveness and scalability of our proposed framework.
****

Metropolis-Hastings Data Augmentation for Graph Neural Networks

Hyeonjin Park, Seunghun Lee, Sihyeon Kim, Jinyoung Park, Jisu Jeong, Kyung-Min Kim, Jung-Woo Ha, Hyunwoo J. Kim

Graph Neural Networks (GNNs) often suffer from weak-generalization due to sparsely labeled data despite their promising results on various graph-based tasks. Da

ta augmentation is a prevalent remedy to improve the generalization ability of models in many domains. However, due to the non-Euclidean nature of data space and the dependencies between samples, designing effective augmentation on graphs is challenging. In this paper, we propose a novel framework Metropolis-Hastings Data Augmentation (MH-Aug) that draws augmented graphs from an explicit target distribution for semi-supervised learning. MH-Aug produces a sequence of augmented graphs from the target distribution enables flexible control of the strength and diversity of augmentation. Since the direct sampling from the complex target distribution is challenging, we adopt the Metropolis-Hastings algorithm to obtain the augmented samples. We also propose a simple and effective semi-supervised learning strategy with generated samples from MH-Aug. Our extensive experiments demonstrate that MH-Aug can generate a sequence of samples according to the target distribution to significantly improve the performance of GNNs.
****

Strategic Behavior is Bliss: Iterative Voting Improves Social Welfare
Joshua Kavner, Lirong Xia
Recent work in iterative voting has defined the additive dynamic price of anarchy (ADPoA) as the difference in social welfare between the truthful and worst-case equilibrium profiles resulting from repeated strategic manipulations. While iterative plurality has been shown to only return alternatives with at most one less initial votes than the truthful winner, it is less understood how agents' welfare changes in equilibrium. To this end, we differentiate agents' utility from their manipulation mechanism and determine iterative plurality's ADPoA in the worst- and average-cases. We first prove that the worst-case ADPoA is linear in the number of agents. To overcome this negative result, we study the average-case ADPoA and prove that equilibrium winners have a constant order welfare advantage over the truthful winner in expectation. Our positive results illustrate the prospect for social welfare to increase due to strategic manipulation.
****

Agnostic Reinforcement Learning with Low-Rank MDPs and Rich Observations
Ayush Sekhari, Christoph Dann, Mehryar Mohri, Yishay Mansour, Karthik Sridharan
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

Functional Regularization for Reinforcement Learning via Learned Fourier Features
Alexander Li, Deepak Pathak
We propose a simple architecture for deep reinforcement learning by embedding inputs into a learned Fourier basis and show that it improves the sample efficiency of both state-based and image-based RL. We perform infinite-width analysis of our architecture using the Neural Tangent Kernel and theoretically show that tuning the initial variance of the Fourier basis is equivalent to functional regularization of the learned deep network. That is, these learned Fourier features allow for adjusting the degree to which networks underfit or overfit different frequencies in the training data, and hence provide a controlled mechanism to improve the stability and performance of RL optimization. Empirically, this allows us to prioritize learning low-frequency functions and speed up learning by reducing networks' susceptibility to noise in the optimization process, such as during Bellman updates. Experiments on standard state-based and image-based RL benchmarks show clear benefits of our architecture over the baselines.
****

Adaptive First-Order Methods Revisited: Convex Minimization without Lipschitz Requirements
Kimon Antonakopoulos, Panayotis Mertikopoulos
We propose a new family of adaptive first-order methods for a class of convex minimization problems that may fail to be Lipschitz continuous or smooth in the standard sense. Specifically, motivated by a recent flurry of activity on non-Lipschitz (NoLips) optimization, we consider problems that are continuous or smooth

relative to a reference Bregman function – as opposed to a global, ambient norm (Euclidean or otherwise). These conditions encompass a wide range ofproblems wit h singular objective, such as Fisher markets, Poisson tomography, D-design, and the like. In this setting, the application of existing order-optimal adaptive me thods – like UnixGrad or AcceleGrad – is not possible, especially in the presenc e of randomness and uncertainty. The proposed method, adaptive mirror descent (A daMir), aims to close this gap by concurrently achieving min-max optimal rates i n problems that are relatively continuous or smooth, including stochastic ones.
****
Adapting to function difficulty and growth conditions in private optimization
Hilal Asi, Daniel Levy, John C. Duchi
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-auth ors prior to requesting a name change in the electronic proceedings.
****
Support Recovery of Sparse Signals from a Mixture of Linear Measurements
Soumyabrata Pal, Arya Mazumdar, Venkata Gandikota
Requests for name changes in the electronic proceedings will be accepted with 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 Gradient Descent-Ascent and Consensus Optimization for Smooth Games: Convergence Analysis under Expected Co-coercivity
Nicolas Loizou, Hugo Berard, Gauthier Gidel, Ioannis Mitliagkas, Simon Lacoste-J ulien
Two of the most prominent algorithms for solving unconstrained smooth games are the classical stochastic gradient descent-ascent (SGDA) and the recently introdu ced stochastic consensus optimization (SCO) [Mescheder et al., 2017]. SGDA is kn own to converge to a stationary point for specific classes of games, but current convergence analyses require a bounded variance assumption. SCO is used success fully for solving large-scale adversarial problems, but its convergence guarante es are limited to its deterministic variant. In this work, we introduce the expe cted co-coercivity condition, explain its benefits, and provide the first last-i terate convergence guarantees of SGDA and SCO under this condition for solving a class of stochastic variational inequality problems that are potentially non-mo notone. We prove linear convergence of both methods to a neighborhood of the sol ution when they use constant step-size, and we propose insightful stepsize-switc hing rules to guarantee convergence to the exact solution. In addition, our conv ergence guarantees hold under the arbitrary sampling paradigm, and as such, we g ive insights into the complexity of minibatching.
****
Tighter Expected Generalization Error Bounds via Wasserstein Distance
Borja Rodríguez Gálvez, German Bassi, Ragnar Thobaben, Mikael Skoglund
Requests for name changes in the electronic proceedings will be accepted with 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.
****
Unifying Width-Reduced Methods for Quasi-Self-Concordant Optimization
Deeksha Adil, Brian Bullins, Sushant Sachdeva
Requests for name changes in the electronic proceedings will be accepted with 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.
****
Bridging the Imitation Gap by Adaptive Insubordination
Luca Weihs, Unnat Jain, Iou-Jen Liu, Jordi Salvador, Svetlana Lazebnik, Aniruddh a Kembhavi, Alex Schwing

In practice, imitation learning is preferred over pure reinforcement learning wh
enever it is possible to design a teaching agent to provide expert supervision.
However, we show that when the teaching agent makes decisions with access to pri
vileged information that is unavailable to the student, this information is marg
inalized during imitation learning, resulting in an "imitation gap" and, potenti
ally, poor results. Prior work bridges this gap via a progression from imitation
 learning to reinforcement learning. While often successful, gradual progression
 fails for tasks that require frequent switches between exploration and memoriza
tion. To better address these tasks and alleviate the imitation gap we propose '
Adaptive Insubordination' (ADVISOR). ADVISOR dynamically weights imitation and r
eward-based reinforcement learning losses during training, enabling on-the-fly s
witching between imitation and exploration. On a suite of challenging tasks set
within gridworlds, multi-agent particle environments, and high-fidelity 3D simul
ators, we show that on-the-fly switching with ADVISOR outperforms pure imitation
, pure reinforcement learning, as well as their sequential and parallel combinat
ions.
****
Adversarial Robustness with Non-uniform Perturbations
Ecenaz Erdemir, Jeffrey Bickford, Luca Melis, Sergul Aydore
Robustness of machine learning models is critical for security related applicati
ons, where real-world adversaries are uniquely focused on evading neural network
 based detectors. Prior work mainly focus on crafting adversarial examples (AEs)
 with small uniform norm-bounded perturbations across features to maintain the r
equirement of imperceptibility. However, uniform perturbations do not result in
realistic AEs in domains such as malware, finance, and social networks. For thes
e types of applications, features typically have some semantically meaningful de
pendencies. The key idea of our proposed approach is to enable non-uniform pertu
rbations that can adequately represent these feature dependencies during adversa
rial training. We propose using characteristics of the empirical data distributi
on, both on correlations between the features and the importance of the features
 themselves. Using experimental datasets for malware classification, credit risk
 prediction, and spam detection, we show that our approach is more robust to rea
l-world attacks. Finally, we present robustness certification utilizing non-unif
orm perturbation bounds, and show that non-uniform bounds achieve better certifi
cation.
****
Container: Context Aggregation Networks
peng gao, Jiasen Lu, hongsheng Li, Roozbeh Mottaghi, Aniruddha Kembhavi
Convolutional neural networks (CNNs) are ubiquitous in computer vision, with a m
yriad of effective and efficient variations. Recently, Transformers -- originall
y introduced in natural language processing -- have been increasingly adopted in
 computer vision. While early adopters continued to employ CNN backbones, the la
test networks are end-to-end CNN-free Transformer solutions. A recent surprising
 finding now shows that a simple MLP based solution without any traditional conv
olutional or Transformer components can produce effective visual representations
. While CNNs, Transformers and MLP-Mixers may be considered as completely dispar
ate architectures, we provide a unified view showing that they are in fact speci
al cases of a more general method to aggregate spatial context in a neural netwo
rk stack. We present the \model (CONText AggregatIon NEtwoRk), a general-purpose
 building block for multi-head context aggregation that can exploit long-range i
nteractions \emph{a la} Transformers while still exploiting the inductive bias o
f the local convolution operation leading to faster convergence speeds, often se
en in CNNs. Our \model architecture achieves 82.7 \% Top-1 accuracy on ImageNet
using 22M parameters, +2.8 improvement compared with DeiT-Small, and can converg
e to 79.9 \% Top-1 accuracy in just 200 epochs. In contrast to Transformer-based
 methods that do not scale well to downstream tasks that rely on larger input im
age resolutions, our efficient network, named \modellight, can be employed in ob
ject detection and instance segmentation networks such as DETR, RetinaNet and Ma
sk-RCNN to obtain an impressive detection mAP of 38.9, 43.8, 45.1 and mask mAP o
f 41.3, providing large improvements of 6.6, 7.3, 6.9 and 6.6 pts respectively,

compared to a ResNet-50 backbone with a comparable compute and parameter size. O
ur method also achieves promising results on self-supervised learning compared t
o DeiT on the DINO framework. Code is released at https://github.com/allenai/con
tainer.
****
ConE: Cone Embeddings for Multi-Hop Reasoning over Knowledge Graphs
Zhanqiu Zhang, Jie Wang, Jiajun Chen, Shuiwang Ji, Feng Wu
Query embedding (QE)---which aims to embed entities and first-order logical (FOL
) queries in low-dimensional spaces---has shown great power in multi-hop reasoni
ng over knowledge graphs. Recently, embedding entities and queries with geometri
c shapes becomes a promising direction, as geometric shapes can naturally repres
ent answer sets of queries and logical relationships among them. However, existi
ng geometry-based models have difficulty in modeling queries with negation, whic
h significantly limits their applicability. To address this challenge, we propos
e a novel query embedding model, namely \textbf{Con}e \textbf{E}mbeddings (ConE)
, which is the first geometry-based QE model that can handle all the FOL operati
ons, including conjunction, disjunction, and negation. Specifically, ConE repres
ents entities and queries as Cartesian products of two-dimensional cones, where
the intersection and union of cones naturally model the conjunction and disjunct
ion operations. By further noticing that the closure of complement of cones rema
ins cones, we design geometric complement operators in the embedding space for t
he negation operations. Experiments demonstrate that ConE significantly outperfo
rms existing state-of-the-art methods on benchmark datasets.
****
Federated Hyperparameter Tuning: Challenges, Baselines, and Connections to Weigh
t-Sharing
Mikhail Khodak, Renbo Tu, Tian Li, Liam Li, Maria-Florina F. Balcan, Virginia Sm
ith, Ameet Talwalkar
Tuning hyperparameters is a crucial but arduous part of the machine learning pip
eline. Hyperparameter optimization is even more challenging in federated learnin
g, where models are learned over a distributed network of heterogeneous devices;
 here, the need to keep data on device and perform local training makes it diffi
cult to efficiently train and evaluate configurations. In this work, we investig
ate the problem of federated hyperparameter tuning. We first identify key challe
nges and show how standard approaches may be adapted to form baselines for the f
ederated setting. Then, by making a novel connection to the neural architecture
search technique of weight-sharing, we introduce a new method, FedEx, to acceler
ate federated hyperparameter tuning that is applicable to widely-used federated
optimization methods such as FedAvg and recent variants. Theoretically, we show
that a FedEx variant correctly tunes the on-device learning rate in the setting
of online convex optimization across devices. Empirically, we show that FedEx ca
n outperform natural baselines for federated hyperparameter tuning by several pe
rcentage points on the Shakespeare, FEMNIST, and CIFAR-10 benchmarks—obtaining h
igher accuracy using the same training budget.
****
Training for the Future: A Simple Gradient Interpolation Loss to Generalize Alon
g Time
Anshul Nasery, Soumyadeep Thakur, Vihari Piratla, Abir De, Sunita Sarawagi
In several real world applications, machine learning models are deployed to make
 predictions on data whose distribution changes gradually along time, leading to
 a drift between the train and test distributions. Such models are often re-trai
ned on new data periodically, and they hence need to generalize to data not too
far into the future. In this context, there is much prior work on enhancing temp
oral generalization, e.g. continuous transportation of past data, kernel smoothe
d time-sensitive parameters and more recently, adversarial learning of time-inva
riant features. However, these methods share several limitations, e.g, poor scal
ability, training instability, and dependence on unlabeled data from the future.
 Responding to the above limitations, we propose a simple method that starts wit
h a model with time-sensitive parameters but regularizes its temporal complexity
 using a Gradient Interpolation (GI) loss. GI allows the decision boundary to c

hange along time and can still prevent overfitting to the limited training time snapshots by allowing task-specific control over changes along time. We compare our method to existing baselines on multiple real-world datasets, which show that GI outperforms more complicated generative and adversarial approaches on the one hand, and simpler gradient regularization methods on the other.
****

Agent Modelling under Partial Observability for Deep Reinforcement Learning
Georgios Papoudakis, Filippos Christianos, Stefano Albrecht
Modelling the behaviours of other agents is essential for understanding how agents interact and making effective decisions. Existing methods for agent modelling commonly assume knowledge of the local observations and chosen actions of the modelled agents during execution. To eliminate this assumption, we extract representations from the local information of the controlled agent using encoder-decoder architectures. Using the observations and actions of the modelled agents during training, our models learn to extract representations about the modelled agents conditioned only on the local observations of the controlled agent. The representations are used to augment the controlled agent's decision policy which is trained via deep reinforcement learning; thus, during execution, the policy does not require access to other agents' information. We provide a comprehensive evaluation and ablations studies in cooperative, competitive and mixed multi-agent environments, showing that our method achieves significantly higher returns than baseline methods which do not use the learned representations.
****

Leveraging Distribution Alignment via Stein Path for Cross-Domain Cold-Start Recommendation
Weiming Liu, Jiajie Su, Chaochao Chen, Xiaolin Zheng
Cross-Domain Recommendation (CDR) has been popularly studied to utilize different domain knowledge to solve the cold-start problem in recommender systems. In this paper, we focus on the Cross-Domain Cold-Start Recommendation (CDCSR) problem. That is, how to leverage the information from a source domain, where items are 'warm', to improve the recommendation performance of a target domain, where items are 'cold'. Unfortunately, previous approaches on cold-start and CDR cannot reduce the latent embedding discrepancy across domains efficiently and lead to model degradation. To address this issue, we propose DisAlign, a cross-domain recommendation framework for the CDCSR problem, which utilizes both rating and auxiliary representations from the source domain to improve the recommendation performance of the target domain. Specifically, we first propose Stein path alignment for aligning the latent embedding distributions across domains, and then further propose its improved version, i.e., proxy Stein path, which can reduce the operation consumption and improve efficiency. Our empirical study on Douban and Amazon datasets demonstrate that DisAlign significantly outperforms the state-of-the-art models under the CDCSR setting.
****

Conservative Offline Distributional Reinforcement Learning
Yecheng Ma, Dinesh Jayaraman, Osbert Bastani
Many reinforcement learning (RL) problems in practice are offline, learning purely from observational data. A key challenge is how to ensure the learned policy is safe, which requires quantifying the risk associated with different actions. In the online setting, distributional RL algorithms do so by learning the distribution over returns (i.e., cumulative rewards) instead of the expected return; beyond quantifying risk, they have also been shown to learn better representations for planning. We proposeConservative Offline Distributional Actor Critic (CODAC), an offline RL algorithm suitable for both risk-neutral and risk-averse domains. CODAC adapts distributional RL to the offline setting by penalizing the predicted quantiles of the return for out-of-distribution actions. We prove that CODAC learns a conservative return distribution---in particular, for finite MDPs, CODAC converges to an uniform lower bound on the quantiles of the return distribution; our proof relies on a novel analysis of the distributional Bellman operator. In our experiments, on two challenging robot navigation tasks, CODAC successfully learns risk-averse policies using offline data collected purely from risk-n

eutral agents. Furthermore, CODAC is state-of-the-art on the D4RL MuJoCo benchmark in terms of both expected and risk-sensitive performance.
****
Separation Results between Fixed-Kernel and Feature-Learning Probability Metrics
Carles Domingo i Enrich, Youssef Mroueh
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Risk Minimization from Adaptively Collected Data: Guarantees for Supervised and Policy Learning
Aurelien Bibaut, Nathan Kallus, Maria Dimakopoulou, Antoine Chambaz, Mark van der Laan
Empirical risk minimization (ERM) is the workhorse of machine learning, whether for classification and regression or for off-policy policy learning, but its model-agnostic guarantees can fail when we use adaptively collected data, such as the result of running a contextual bandit algorithm. We study a generic importance sampling weighted ERM algorithm for using adaptively collected data to minimize the average of a loss function over a hypothesis class and provide first-of-their-kind generalization guarantees and fast convergence rates. Our results are based on a new maximal inequality that carefully leverages the importance sampling structure to obtain rates with the good dependence on the exploration rate in the data. For regression, we provide fast rates that leverage the strong convexity of squared-error loss. For policy learning, we provide regret guarantees that close an open gap in the existing literature whenever exploration decays to zero, as is the case for bandit-collected data. An empirical investigation validates our theory.
****
Bayesian Optimization with High-Dimensional Outputs
Wesley J. Maddox, Maximilian Balandat, Andrew G. Wilson, Eytan Bakshy
Bayesian optimization is a sample-efficient black-box optimization procedure that is typically applied to a small number of independent objectives. However, in practice we often wish to optimize objectives defined over many correlated outcomes (or “tasks”). For example, scientists may want to optimize the coverage of a cell tower network across a dense grid of locations. Similarly, engineers may seek to balance the performance of a robot across dozens of different environments via constrained or robust optimization. However, the Gaussian Process (GP) models typically used as probabilistic surrogates for multi-task Bayesian optimization scale poorly with the number of outcomes, greatly limiting applicability. We devise an efficient technique for exact multi-task GP sampling that combines exploiting Kronecker structure in the covariance matrices with Matheron’s identity, allowing us to perform Bayesian optimization using exact multi-task GP models with tens of thousands of correlated outputs. In doing so, we achieve substantial improvements in sample efficiency compared to existing approaches that model solely the outcome metrics. We demonstrate how this unlocks a new class of applications for Bayesian optimization across a range of tasks in science and engineering, including optimizing interference patterns of an optical interferometer with 65,000 outputs.
****
Finding Optimal Tangent Points for Reducing Distortions of Hard-label Attacks
Chen Ma, Xiangyu Guo, Li Chen, Jun-Hai Yong, Yisen 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.
****
Scalable Diverse Model Selection for Accessible Transfer Learning
Daniel Bolya, Rohit Mittapalli, Judy Hoffman
With the preponderance of pretrained deep learning models available off-the-shel

f from model banks today, finding the best weights to fine-tune to your use-case
 can be a daunting task. Several methods have recently been proposed to find goo
d models for transfer learning, but they either don't scale well to large model
banks or don't perform well on the diversity of off-the-shelf models. Ideally th
e question we want to answer is, "given some data and a source model, can you qu
ickly predict the model's accuracy after fine-tuning?" In this paper, we formali
ze this setting as "Scalable Diverse Model Selection" and propose several benchm
arks for evaluating on this task. We find that existing model selection and tran
sferability estimation methods perform poorly here and analyze why this is the c
ase. We then introduce simple techniques to improve the performance and speed of
 these algorithms. Finally, we iterate on existing methods to create PARC, which
 outperforms all other methods on diverse model selection. We have released the
benchmarks and method code in hope to inspire future work in model selection for
 accessible transfer learning.
****
Light Field Networks: Neural Scene Representations with Single-Evaluation Render
ing
Vincent Sitzmann, Semon Rezchikov, Bill Freeman, Josh Tenenbaum, Fredo Durand
Inferring representations of 3D scenes from 2D observations is a fundamental pro
blem of computer graphics, computer vision, and artificial intelligence. Emergin
g 3D-structured neural scene representations are a promising approach to 3D scen
e understanding. In this work, we propose a novel neural scene representation, L
ight Field Networks or LFNs, which represent both geometry and appearance of the
 underlying 3D scene in a 360-degree, four-dimensional light field parameterized
 via a neural implicit representation. Rendering a ray from an LFN requires onl
y a single network evaluation, as opposed to hundreds of evaluations per ray for
 ray-marching or volumetric based renderers in 3D-structured neural scene repres
entations. In the setting of simple scenes, we leverage meta-learning to learn
a prior over LFNs that enables multi-view consistent light field reconstruction
from as little as a single image observation. This results in dramatic reduction
s in time and memory complexity, and enables real-time rendering. The cost of st
oring a 360-degree light field via an LFN is two orders of magnitude lower than
conventional methods such as the Lumigraph. Utilizing the analytical differenti
ability of neural implicit representations and a novel parameterization of light
 space, we further demonstrate the extraction of sparse depth maps from LFNs.
****
ViSER: Video-Specific Surface Embeddings for Articulated 3D Shape Reconstruction
Gengshan Yang, Deqing Sun, Varun Jampani, Daniel Vlasic, Forrester Cole, Ce Liu,
 Deva Ramanan
We introduce ViSER, a method for recovering articulated 3D shapes and dense3D tr
ajectories from monocular videos. Previous work on high-quality reconstruction
of dynamic 3D shapes typically relies on multiple camera views, strong category-
specific priors, or 2D keypoint supervision. We show that none of these are requ
ired if one can reliably estimate long-range correspondences in a video, making
use of only 2D object masks and two-frame optical flow as inputs. ViSER infers c
orrespondences by matching 2D pixels to a canonical, deformable 3D mesh via vid
eo-specific surface embeddings that capture the pixel appearance of each surface
 point. These embeddings behave as a continuous set of keypoint descriptors def
ined over the mesh surface, which can be used to establish dense long-range corr
espondences across pixels. The surface embeddings are implemented as coordinate
-based MLPs that are fit to each video via self-supervised losses.Experimental r
esults show that ViSER compares favorably against prior work on challenging vide
os of humans with loose clothing and unusual poses as well as animals videos fro
m DAVIS and YTVOS. Project page: viser-shape.github.io.
****
Understanding the Effect of Stochasticity in Policy Optimization
Jincheng Mei, Bo Dai, Chenjun Xiao, Csaba Szepesvari, Dale Schuurmans
Requests for name changes in the electronic proceedings will be accepted with 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.
****
Fine-Grained Zero-Shot Learning with DNA as Side Information
Sarkhan Badirli, Zeynep Akata, George Mohler, Christine Picard, Mehmet M Dundar
Fine-grained zero-shot learning task requires some form of side-information to t
ransfer discriminative information from seen to unseen classes. As manually anno
tated visual attributes are extremely costly and often impractical to obtain for
 a large number of classes, in this study we use DNA as a side information for t
he first time for fine-grained zero-shot classification of species. Mitochondria
l DNA plays an important role as a genetic marker in evolutionary biology and ha
s been used to achieve near perfect accuracy in species classification of living
 organisms. We implement a simple hierarchical Bayesian model that uses DNA info
rmation to establish the hierarchy in the image space and employs local priors t
o define surrogate classes for unseen ones. On the benchmark CUB dataset we show
 that DNA can be equally promising, yet in general a more accessible alternative
 than word vectors as a side information. This is especially important as obtain
ing robust word representations for fine-grained species names is not a practica
ble goal when information about these species in free-form text is limited. On a
 newly compiled fine-grained insect dataset that uses DNA information from over
a thousand species we show that the Bayesian approach outperforms state-of-the-a
rt by a wide margin.
****
Optimal Underdamped Langevin MCMC Method
Zhengmian Hu, Feihu Huang, Heng Huang
Requests for name changes in the electronic proceedings will be accepted with no
 questions asked. However name changes may cause bibliographic tracking issues.
  Authors are asked to consider this carefully and discuss it with their co-auth
ors prior to requesting a name change in the electronic proceedings.
****
Scheduling jobs with stochastic holding costs
Dabeen Lee, Milan Vojnovic
Requests for name changes in the electronic proceedings will be accepted with 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.
****
REMIPS: Physically Consistent 3D Reconstruction of Multiple Interacting People u
nder Weak Supervision
Mihai Fieraru, Mihai Zanfir, Teodor Szente, Eduard Bazavan, Vlad Olaru, Cristian
 Sminchisescu
The three-dimensional reconstruction of multiple interacting humans given a mono
cular image is crucial for the general task of scene understanding, as capturing
 the subtleties of interaction is often the very reason for taking a picture. Cu
rrent 3D human reconstruction methods either treat each person independently, ig
noring most of the context, or reconstruct people jointly, but cannot recover in
teractions correctly when people are in close proximity. In this work, we introd
uce \textbf{REMIPS}, a model for 3D \underline{Re}construction of \underline{M}u
ltiple \underline{I}nteracting \underline{P}eople under Weak \underline{S}upervi
sion. \textbf{REMIPS} can reconstruct a variable number of people directly from
monocular images. At the core of our methodology stands a novel transformer netw
ork that combines unordered person tokens (one for each detected human) with pos
itional-encoded tokens from image features patches. We introduce a novel unified
 model for self- and interpenetration-collisions based on a mesh approximation c
omputed by applying decimation operators. We rely on self-supervised losses for
flexibility and generalisation in-the-wild and incorporate self-contact and inte
raction-contact losses directly into the learning process. With \textbf{REMIPS},
 we report state-of-the-art quantitative results on common benchmarks even in ca
ses where no 3D supervision is used. Additionally, qualitative visual results sh
ow that our reconstructions are plausible in terms of pose and shape and coheren
t for challenging images, collected in-the-wild, where people are often interact

ing.
****
Differentiable Annealed Importance Sampling and the Perils of Gradient Noise
Guodong Zhang, Kyle Hsu, Jianing Li, Chelsea Finn, Roger B. Grosse
Annealed importance sampling (AIS) and related algorithms are highly effective tools for marginal likelihood estimation, but are not fully differentiable due to the use of Metropolis-Hastings correction steps. Differentiability is a desirable property as it would admit the possibility of optimizing marginal likelihood as an objective using gradient-based methods. To this end, we propose Differentiable AIS (DAIS), a variant of AIS which ensures differentiability by abandoning the Metropolis-Hastings corrections. As a further advantage, DAIS allows for mini-batch gradients. We provide a detailed convergence analysis for Bayesian linear regression which goes beyond previous analyses by explicitly accounting for the sampler not having reached equilibrium. Using this analysis, we prove that DAIS is consistent in the full-batch setting and provide a sublinear convergence rate. Furthermore, motivated by the problem of learning from large-scale datasets, we study a stochastic variant of DAIS that uses mini-batch gradients. Surprisingly, stochastic DAIS can be arbitrarily bad due to a fundamental incompatibility between the goals of last-iterate convergence to the posterior and elimination of the accumulated stochastic error. This is in stark contrast with other settings such as gradient-based optimization and Langevin dynamics, where the effect of gradient noise can be washed out by taking smaller steps. This indicates that annealing-based marginal likelihood estimation with stochastic gradients may require new ideas.
****
PSD Representations for Effective Probability Models
Alessandro Rudi, Carlo Ciliberto
Finding a good way to model probability densities is key to probabilistic inference. An ideal model should be able to concisely approximate any probability while being also compatible with two main operations: multiplications of two models (product rule) and marginalization with respect to a subset of the random variables (sum rule). In this work, we show that a recently proposed class of positive semi-definite (PSD) models for non-negative functions is particularly suited to this end. In particular, we characterize both approximation and generalization capabilities of PSD models, showing that they enjoy strong theoretical guarantees. Moreover, we show that we can perform efficiently both sum and product rule in closed form via matrix operations, enjoying the same versatility of mixture models. Our results open the way to applications of PSD models to density estimation, decision theory, and inference.
****
Exploiting a Zoo of Checkpoints for Unseen Tasks
Jiaji Huang, Qiang Qiu, Kenneth Church
There are so many models in the literature that it is difficult for practitioners to decide which combinations are likely to be effective for a new task. This paper attempts to address this question by capturing relationships among checkpoints published on the web. We model the space of tasks as a Gaussian process. The covariance can be estimated from checkpoints and unlabeled probing data. With the Gaussian process, we can identify representative checkpoints by a maximum mutual information criterion. This objective is submodular. A greedy method identifies representatives that are likely to "cover'' the task space. These representatives generalize to new tasks with superior performance. Empirical evidence is provided for applications from both computational linguistics as well as computer vision.
****
Towards Open-World Feature Extrapolation: An Inductive Graph Learning Approach
Qitian Wu, Chenxiao Yang, Junchi Yan
We target open-world feature extrapolation problem where the feature space of input data goes through expansion and a model trained on partially observed features needs to handle new features in test data without further retraining. The problem is of much significance for dealing with features incrementally collected f

rom different fields. To this end, we propose a new learning paradigm with graph representation and learning. Our framework contains two modules: 1) a backbone network (e.g., feedforward neural nets) as a lower model takes features as input and outputs predicted labels; 2) a graph neural network as an upper model learns to extrapolate embeddings for new features via message passing over a feature-data graph built from observed data. Based on our framework, we design two training strategies, a self-supervised approach and an inductive learning approach, to endow the model with extrapolation ability and alleviate feature-level over-fitting. We also provide theoretical analysis on the generalization error on test data with new features, which dissects the impact of training features and algorithms on generalization performance. Our experiments over several classification datasets and large-scale advertisement click prediction datasets demonstrate that our model can produce effective embeddings for unseen features and significantly outperforms baseline methods that adopt KNN and local aggregation.
****

Adversarial Teacher-Student Representation Learning for Domain Generalization
Fu-En Yang, Yuan-Chia Cheng, Zu-Yun Shiau, Yu-Chiang Frank Wang
Domain generalization (DG) aims to transfer the learning task from a single or multiple source domains to unseen target domains. To extract and leverage the information which exhibits sufficient generalization ability, we propose a simple yet effective approach of Adversarial Teacher-Student Representation Learning, with the goal of deriving the domain generalizable representations via generating and exploring out-of-source data distributions. Our proposed framework advances Teacher-Student learning in an adversarial learning manner, which alternates between knowledge-distillation based representation learning and novel-domain data augmentation. The former progressively updates the teacher network for deriving domain-generalizable representations, while the latter synthesizes data out-of-source yet plausible distributions. Extensive image classification experiments on benchmark datasets in multiple and single source DG settings confirm that, our model exhibits sufficient generalization ability and performs favorably against state-of-the-art DG methods.
****

Stochastic bandits with groups of similar arms.
Fabien Pesquerel, Hassan SABER, Odalric-Ambrym Maillard
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

Tracking Without Re-recognition in Humans and Machines
Drew Linsley, Girik Malik, Junkyung Kim, Lakshmi Narasimhan Govindarajan, Ennio Mingolla, Thomas Serre
Imagine trying to track one particular fruitfly in a swarm of hundreds. Higher biological visual systems have evolved to track moving objects by relying on both their appearance and their motion trajectories. We investigate if state-of-the-art spatiotemporal deep neural networks are capable of the same. For this, we introduce PathTracker, a synthetic visual challenge that asks human observers and machines to track a target object in the midst of identical-looking "distractor" objects. While humans effortlessly learn PathTracker and generalize to systematic variations in task design, deep networks struggle. To address this limitation, we identify and model circuit mechanisms in biological brains that are implicated in tracking objects based on motion cues. When instantiated as a recurrent network, our circuit model learns to solve PathTracker with a robust visual strategy that rivals human performance and explains a significant proportion of their decision-making on the challenge. We also show that the success of this circuit model extends to object tracking in natural videos. Adding it to a transformer-based architecture for object tracking builds tolerance to visual nuisances that affect object appearance, establishing the new state of the art on the large-scale TrackingNet challenge. Our work highlights the importance of understanding human vision to improve computer vision.

****
Rethinking conditional GAN training: An approach using geometrically structured latent manifolds
Sameera Ramasinghe, Moshiur Farazi, Salman H Khan, Nick Barnes, Stephen Gould
Conditional GANs (cGAN), in their rudimentary form, suffer from critical drawbacks such as the lack of diversity in generated outputs and distortion between the latent and output manifolds. Although efforts have been made to improve results, they can suffer from unpleasant side-effects such as the topology mismatch between latent and output spaces. In contrast, we tackle this problem from a geometrical perspective and propose a novel training mechanism that increases both the diversity and the visual quality of a vanilla cGAN, by systematically encouraging a bi-lipschitz mapping between the latent and the output manifolds. We validate the efficacy of our solution on a baseline cGAN (i.e., Pix2Pix) which lacks diversity, and show that by only modifying its training mechanism (i.e., with our proposed Pix2Pix-Geo), one can achieve more diverse and realistic outputs on a broad set of image-to-image translation tasks.
****
How to transfer algorithmic reasoning knowledge to learn new algorithms?
Louis-Pascal Xhonneux, Andreea-Ioana Deac, Petar Veli■kovi■, Jian Tang
Learning to execute algorithms is a fundamental problem that has been widely studied. Prior work (Veli■kovi■ et al., 2019) has shown that to enable systematic generalisation on graph algorithms it is critical to have access to the intermediate steps of the program/algorithm. In many reasoning tasks, where algorithmic-style reasoning is important, we only have access to the input and output examples. Thus, inspired by the success of pre-training on similar tasks or data in Natural Language Processing (NLP) and Computer vision, we set out to study how we can transfer algorithmic reasoning knowledge. Specifically, we investigate how we can use algorithms for which we have access to the execution trace to learn to solve similar tasks for which we do not. We investigate two major classes of graph algorithms, parallel algorithms such as breadth-first search and Bellman-Ford and sequential greedy algorithms such as Prims and Dijkstra. Due to the fundamental differences between algorithmic reasoning knowledge and feature extractors such as used in Computer vision or NLP, we hypothesis that standard transfer techniques will not be sufficient to achieve systematic generalisation. To investigate this empirically we create a dataset including 9 algorithms and 3 different graph types. We validate this empirically and show how instead multi-task learning can be used to achieve the transfer of algorithmic reasoning knowledge.
****
Fast Axiomatic Attribution for Neural Networks
Robin Hesse, Simone Schaub-Meyer, Stefan Roth
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
OSOA: One-Shot Online Adaptation of Deep Generative Models for Lossless Compression
Chen Zhang, Shifeng Zhang, Fabio Maria Carlucci, Zhenguo 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.
****
Compressive Visual Representations
Kuang-Huei Lee, Anurag Arnab, Sergio Guadarrama, John Canny, Ian Fischer
Learning effective visual representations that generalize well without human supervision is a fundamental problem in order to apply Machine Learning to a wide variety of tasks. Recently, two families of self-supervised methods, contrastive learning and latent bootstrapping, exemplified by SimCLR and BYOL respectively, have made significant progress. In this work, we hypothesize that adding explici

t information compression to these algorithms yields better and more robust repr esentations. We verify this by developing SimCLR and BYOL formulations compatibl e with the Conditional Entropy Bottleneck (CEB) objective, allowing us to both m easure and control the amount of compression in the learned representation, and observe their impact on downstream tasks. Furthermore, we explore the relationsh ip between Lipschitz continuity and compression, showing a tractable lower bound on the Lipschitz constant of the encoders we learn. As Lipschitz continuity is closely related to robustness, this provides a new explanation for why compresse d models are more robust. Our experiments confirm that adding compression to Sim CLR and BYOL significantly improves linear evaluation accuracies and model robus tness across a wide range of domain shifts. In particular, the compressed versio n of BYOL achieves 76.0% Top-1 linear evaluation accuracy on ImageNet with ResNe t-50, and 78.8% with ResNet-50 2x.
****
Multi-Armed Bandits with Bounded Arm-Memory: Near-Optimal Guarantees for Best-Ar m Identification and Regret Minimization
Arnab Maiti, Vishakha Patil, Arindam Khan
Requests for name changes in the electronic proceedings will be accepted with 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.
****
Grounding inductive biases in natural images: invariance stems from variations i n data
Diane Bouchacourt, Mark Ibrahim, Ari Morcos
To perform well on unseen and potentially out-of-distribution samples, it is des irable for machine learning models to have a predictable response with respect t o transformations affecting the factors of variation of the input. Here, we stud y the relative importance of several types of inductive biases towards such pred ictable behavior: the choice of data, their augmentations, and model architectur es. Invariance is commonly achieved through hand-engineered data augmentation, b ut do standard data augmentations address transformations that explain variation s in real data? While prior work has focused on synthetic data, we attempt here to characterize the factors of variation in a real dataset, ImageNet, and study the invariance of both standard residual networks and the recently proposed visi on transformer with respect to changes in these factors. We show standard augmen tation relies on a precise combination of translation and scale, with translatio n recapturing most of the performance improvement---despite the (approximate) tr anslation invariance built in to convolutional architectures, such as residual n etworks. In fact, we found that scale and translation invariance was similar acr oss residual networks and vision transformer models despite their markedly diffe rent architectural inductive biases. We show the training data itself is the mai n source of invariance, and that data augmentation only further increases the le arned invariances. Notably, the invariances learned during training align with t he ImageNet factors of variation we found. Finally, we find that the main factor s of variation in ImageNet mostly relate to appearance and are specific to each class.
****
Directed Graph Contrastive Learning
Zekun Tong, Yuxuan Liang, Henghui Ding, Yongxing Dai, Xinke Li, Changhu Wang
Graph Contrastive Learning (GCL) has emerged to learn generalizable representati ons from contrastive views. However, it is still in its infancy with two concern s: 1) changing the graph structure through data augmentation to generate contras tive views may mislead the message passing scheme, as such graph changing action deprives the intrinsic graph structural information, especially the directional structure in directed graphs; 2) since GCL usually uses predefined contrastive views with hand-picking parameters, it does not take full advantage of the contr astive information provided by data augmentation, resulting in incomplete struct ure information for models learning. In this paper, we design a directed graph d ata augmentation method called Laplacian perturbation and theoretically analyze

how it provides contrastive information without changing the directed graph stru cture. Moreover, we present a directed graph contrastive learning framework, whi ch dynamically learns from all possible contrastive views generated by Laplacian perturbation. Then we train it using multi-task curriculum learning to progress ively learn from multiple easy-to-difficult contrastive views. We empirically sh ow that our model can retain more structural features of directed graphs than ot her GCL models because of its ability to provide complete contrastive informatio n. Experiments on various benchmarks reveal our dominance over the state-of-the-art approaches.
****
Space-time Mixing Attention for Video Transformer
Adrian Bulat, Juan Manuel Perez Rua, Swathikiran Sudhakaran, Brais Martinez, Geo rgios Tzimiropoulos
This paper is on video recognition using Transformers. Very recent attempts in t his area have demonstrated promising results in terms of recognition accuracy, y et they have been also shown to induce, in many cases, significant computational overheads due to the additional modelling of the temporal information. In this work, we propose a Video Transformer model the complexity of which scales linear ly with the number of frames in the video sequence and hence induces no overhead compared to an image-based Transformer model. To achieve this, our model makes two approximations to the full space-time attention used in Video Transformers: (a) It restricts time attention to a local temporal window and capitalizes on th e Transformer's depth to obtain full temporal coverage of the video sequence. (b) It uses efficient space-time mixing to attend jointly spatial and temporal loc ations without inducing any additional cost on top of a spatial-only attention m odel. We also show how to integrate 2 very lightweight mechanisms for global tem poral-only attention which provide additional accuracy improvements at minimal c omputational cost. We demonstrate that our model produces very high recognition accuracy on the most popular video recognition datasets while at the same time b eing significantly more efficient than other Video Transformer models.
****
Particle Dual Averaging: Optimization of Mean Field Neural Network with Global C onvergence Rate Analysis
Atsushi Nitanda, Denny Wu, Taiji Suzuki
We propose the particle dual averaging (PDA) method, which generalizes the dual averaging method in convex optimization to the optimization over probability dis tributions with quantitative runtime guarantee. The algorithm consists of an inn er loop and outer loop: the inner loop utilizes the Langevin algorithm to approx imately solve for a stationary distribution, which is then optimized in the oute r loop. The method can be interpreted as an extension of the Langevin algorithm to naturally handle nonlinear functional on the probability space. An important application of the proposed method is the optimization of neural network in the mean field regime, which is theoretically attractive due to the presence of nonl inear feature learning, but quantitative convergence rate can be challenging to obtain. By adapting finite-dimensional convex optimization theory into the space of measures, we not only establish global convergence of PDA for two-layer mean field neural networks under more general settings and simpler analysis, but als o provide quantitative polynomial runtime guarantee. Our theoretical results are supported by numerical simulations on neural networks with reasonable size.
****
Learning Tree Interpretation from Object Representation for Deep Reinforcement L earning
Guiliang Liu, Xiangyu Sun, Oliver Schulte, Pascal Poupart
Interpreting Deep Reinforcement Learning (DRL) models is important to enhance tr ust and comply with transparency regulations. Existing methods typically explain a DRL model by visualizing the importance of low-level input features with supe r-pixels, attentions, or saliency maps. Our approach provides an interpretation based on high-level latent object features derived from a disentangled represent ation. We propose a Represent And Mimic (RAMi) framework for training 1) an iden tifiable latent representation to capture the independent factors of variation f

or the objects and 2) a mimic tree that extracts the causal impact of the latent
 features on DRL action values. To jointly optimize both the fidelity and the si
mplicity of a mimic tree, we derive a novel Minimum Description Length (MDL) obj
ective based on the Information Bottleneck (IB) principle. Based on this objecti
ve, we describe a Monte Carlo Regression Tree Search (MCRTS) algorithm that expl
ores different splits to find the IB-optimal mimic tree. Experiments show that o
ur mimic tree achieves strong approximation performance with significantly fewer
 nodes than baseline models. We demonstrate the interpretability of our mimic tr
ee by showing latent traversals, decision rules, causal impacts, and human evalu
ation results.
****
Only Train Once: A One-Shot Neural Network Training And Pruning Framework
Tianyi Chen, Bo Ji, Tianyu Ding, Biyi Fang, Guanyi Wang, Zhihui Zhu, Luming Lian
g, Yixin Shi, Sheng Yi, Xiao Tu
Structured pruning is a commonly used technique in deploying deep neural network
s (DNNs) onto resource-constrained devices. However, the existing pruning method
s are usually heuristic, task-specified, and require an extra fine-tuning proced
ure. To overcome these limitations, we propose a framework that compresses DNNs
into slimmer architectures with competitive performances and significant FLOPs r
eductions by Only-Train-Once (OTO). OTO contains two key steps: (i) we partition
 the parameters of DNNs into zero-invariant groups, enabling us to prune zero gr
oups without affecting the output; and (ii) to promote zero groups, we then form
ulate a structured-sparsity optimization problem, and propose a novel optimizati
on algorithm, Half-Space Stochastic Projected Gradient (HSPG), to solve it, whic
h outperforms the standard proximal methods on group sparsity exploration, and m
aintains comparable convergence. To demonstrate the effectiveness of OTO, we tra
in and compress full models simultaneously from scratch without fine-tuning for
inference speedup and parameter reduction, and achieve state-of-the-art results
on VGG16 for CIFAR10, ResNet50 for CIFAR10 and Bert for SQuAD and competitive re
sult on ResNet50 for ImageNet. The source code is available at https://github.co
m/tianyic/onlytrainonce.
****
Referring Transformer: A One-step Approach to Multi-task Visual Grounding
Muchen Li, Leonid Sigal
As an important step towards visual reasoning, visual grounding (e.g., phrase lo
calization, referring expression comprehension / segmentation) has been widely e
xplored. Previous approaches to referring expression comprehension (REC) or segm
entation (RES) either suffer from limited performance, due to a two-stage setup,
 or require the designing of complex task-specific one-stage architectures. In t
his paper, we propose a simple one-stage multi-task framework for visual groundi
ng tasks. Specifically, we leverage a transformer architecture, where two modali
ties are fused in a visual-lingual encoder. In the decoder, the model learns to
generate contextualized lingual queries which are then decoded and used to direc
tly regress the bounding box and produce a segmentation mask for the correspondi
ng referred regions. With this simple but highly contextualized model, we outper
form state-of-the-art methods by a large margin on both REC and RES tasks. We al
so show that a simple pre-training schedule (on an external dataset) further imp
roves the performance. Extensive experiments and ablations illustrate that our m
odel benefits greatly from contextualized information and multi-task training.
****
Decoupling the Depth and Scope of Graph Neural Networks
Hanqing Zeng, Muhan Zhang, Yinglong Xia, Ajitesh Srivastava, Andrey Malevich, Ra
jgopal Kannan, Viktor Prasanna, Long Jin, Ren Chen
State-of-the-art Graph Neural Networks (GNNs) have limited scalability with resp
ect to the graph and model sizes. On large graphs, increasing the model depth of
ten means exponential expansion of the scope (i.e., receptive field). Beyond jus
t a few layers, two fundamental challenges emerge: 1. degraded expressivity due
 to oversmoothing, and 2. expensive computation due to neighborhood explosion. W
e propose a design principle to decouple the depth and scope of GNNs – to genera
te representation of a target entity (i.e., a node or an edge), we first extract

a localized subgraph as the bounded-size scope, and then apply a GNN of arbitrary depth on top of the subgraph. A properly extracted subgraph consists of a small number of critical neighbors, while excluding irrelevant ones. The GNN, no matter how deep it is, smooths the local neighborhood into informative representation rather than oversmoothing the global graph into “white noise”. Theoretically, decoupling improves the GNN expressive power from the perspectives of graph signal processing (GCN), function approximation (GraphSAGE) and topological learning (GIN). Empirically, on seven graphs (with up to 110M nodes) and six backbone GNN architectures, our design achieves significant accuracy improvement with orders of magnitude reduction in computation and hardware cost.
****
Fast and Memory Efficient Differentially Private-SGD via JL Projections
Zhiqi Bu, Sivakanth Gopi, Janardhan Kulkarni, Yin Tat Lee, Hanwen Shen, Uthaipon Tantipongpipat
Differentially Private-SGD (DP-SGD) of Abadi et al. and its variations are the only known algorithms for private training of large scale neural networks. This algorithm requires computation of per-sample gradients norms which is extremely slow and memory intensive in practice. In this paper, we present a new framework to design differentially private optimizers called DP-SGD-JL and DP-Adam-JL. Our approach uses Johnson–Lindenstrauss (JL) projections to quickly approximate the per-sample gradient norms without exactly computing them, thus making the training time and memory requirements of our optimizers closer to that of their non-DP versions. Unlike previous attempts to make DP-SGD faster which work only on a subset of network architectures or use compiler techniques, we propose an algorithmic solution which works for any network in a black-box manner which is the main contribution of this paper. To illustrate this, on IMDb dataset, we train a Recurrent Neural Network (RNN) to achieve good privacy-vs-accuracy tradeoff, while being significantly faster than DP-SGD and with a similar memory footprint as non-private SGD.
****
Formalizing Generalization and Adversarial Robustness of Neural Networks to Weight Perturbations
Yu-Lin Tsai, Chia-Yi Hsu, Chia-Mu Yu, Pin-Yu Chen
Studying the sensitivity of weight perturbation in neural networks and its impacts on model performance, including generalization and robustness, is an active research topic due to its implications on a wide range of machine learning tasks such as model compression, generalization gap assessment, and adversarial attacks. In this paper, we provide the first integral study and analysis for feed-forward neural networks in terms of the robustness in pairwise class margin and its generalization behavior under weight perturbation. We further design a new theory-driven loss function for training generalizable and robust neural networks against weight perturbations. Empirical experiments are conducted to validate our theoretical analysis. Our results offer fundamental insights for characterizing the generalization and robustness of neural networks against weight perturbations.
****
Pipeline Combinators for Gradual AutoML
Guillaume Baudart, Martin Hirzel, Kiran Kate, Parikshit Ram, Avi Shinnar, Jason Tsay
Automated machine learning (AutoML) can make data scientists more productive. But if machine learning is totally automated, that leaves no room for data scientists to apply their intuition. Hence, data scientists often prefer not total but gradual automation, where they control certain choices and AutoML explores the rest. Unfortunately, gradual AutoML is cumbersome with state-of-the-art tools, requiring large non-compositional code changes. More concise compositional code can be achieved with combinators, a powerful concept from functional programming. This paper introduces a small set of orthogonal combinators for composing machine-learning operators into pipelines. It describes a translation scheme from pipelines and associated hyperparameter schemas to search spaces for AutoML optimizers. On that foundation, this paper presents Lale, an open-source sklearn-

compatible AutoML library, and evaluates it with a user study.
****
Boost Neural Networks by Checkpoints
Feng Wang, Guoyizhe Wei, Qiao Liu, Jinxiang Ou, xian wei, Hairong Lv
Training multiple deep neural networks (DNNs) and averaging their outputs is a simple way to improve the predictive performance. Nevertheless, the multiplied training cost prevents this ensemble method to be practical and efficient. Several recent works attempt to save and ensemble the checkpoints of DNNs, which only requires the same computational cost as training a single network. However, these methods suffer from either marginal accuracy improvements due to the low diversity of checkpoints or high risk of divergence due to the cyclical learning rates they adopted. In this paper, we propose a novel method to ensemble the checkpoints, where a boosting scheme is utilized to accelerate model convergence and maximize the checkpoint diversity. We theoretically prove that it converges by reducing exponential loss. The empirical evaluation also indicates our proposed ensemble outperforms single model and existing ensembles in terms of accuracy and efficiency. With the same training budget, our method achieves 4.16% lower error on Cifar-100 and 6.96% on Tiny-ImageNet with ResNet-110 architecture. Moreover, the adaptive sample weights in our method make it an effective solution to address the imbalanced class distribution. In the experiments, it yields up to 5.02% higher accuracy over single EfficientNet-B0 on the imbalanced datasets.
****
Model Selection for Bayesian Autoencoders
Ba-Hien Tran, Simone Rossi, Dimitrios Milios, Pietro Michiardi, Edwin V. Bonilla, Maurizio Filippone
We develop a novel method for carrying out model selection for Bayesian autoencoders (BAEs) by means of prior hyper-parameter optimization. Inspired by the common practice of type-II maximum likelihood optimization and its equivalence to Kullback-Leibler divergence minimization, we propose to optimize the distributional sliced-Wasserstein distance (DSWD) between the output of the autoencoder and the empirical data distribution. The advantages of this formulation are that we can estimate the DSWD based on samples and handle high-dimensional problems. We carry out posterior estimation of the BAE parameters via stochastic gradient Hamiltonian Monte Carlo and turn our BAE into a generative model by fitting a flexible Dirichlet mixture model in the latent space. Thanks to this approach, we obtain a powerful alternative to variational autoencoders, which are the preferred choice in modern application of autoencoders for representation learning with uncertainty. We evaluate our approach qualitatively and quantitatively using a vast experimental campaign on a number of unsupervised learning tasks and show that, in small-data regimes where priors matter, our approach provides state-of-the-art results, outperforming multiple competitive baselines.
****
Three Operator Splitting with Subgradients, Stochastic Gradients, and Adaptive Learning Rates
Alp Yurtsever, Alex Gu, Suvrit Sra
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Knowledge-Adaptation Priors
Mohammad Emtiyaz E. Khan, Siddharth Swaroop
Humans and animals have a natural ability to quickly adapt to their surroundings, but machine-learning models, when subjected to changes, often require a complete retraining from scratch. We present Knowledge-adaptation priors (K-priors) to reduce the cost of retraining by enabling quick and accurate adaptation for a wide-variety of tasks and models. This is made possible by a combination of weight and function-space priors to reconstruct the gradients of the past, which recovers and generalizes many existing, but seemingly-unrelated, adaptation strategies. Training with simple first-order gradient methods can often recover the exac

t retrained model to an arbitrary accuracy by choosing a sufficiently large memory of the past data. Empirical results show that adaptation with K-priors achieves performance similar to full retraining, but only requires training on a handful of past examples.
****
Provably efficient multi-task reinforcement learning with model transfer
Chicheng Zhang, Zhi Wang
We study multi-task reinforcement learning (RL) in tabular episodic Markov decision processes (MDPs). We formulate a heterogeneous multi-player RL problem, in which a group of players concurrently face similar but not necessarily identical MDPs, with a goal of improving their collective performance through inter-player information sharing. We design and analyze a model-based algorithm, and provide gap-dependent and gap-independent regret upper and lower bounds that characterize the intrinsic complexity of the problem.
****
Predicting Molecular Conformation via Dynamic Graph Score Matching
Shitong Luo, Chence Shi, Minkai Xu, Jian Tang
Predicting stable 3D conformations from 2D molecular graphs has been a long-standing challenge in computational chemistry. Recently, machine learning approaches have demonstrated very promising results compared to traditional experimental and physics-based simulation methods. These approaches mainly focus on modeling the local interactions between neighboring atoms on the molecular graphs and overlook the long-range interactions between non-bonded atoms. However, these non-bonded atoms may be proximal to each other in 3D space, and modeling their interactions is of crucial importance to accurately determine molecular conformations, especially for large molecules and multi-molecular complexes. In this paper, we propose a new approach called Dynamic Graph Score Matching (DGSM) for molecular conformation prediction, which models both the local and long-range interactions by dynamically constructing graph structures between atoms according to their spatial proximity during both training and inference. Specifically, the DGSM directly estimates the gradient fields of the logarithm density of atomic coordinates according to the dynamically constructed graphs using score matching methods. The whole framework can be efficiently trained in an end-to-end fashion. Experiments across multiple tasks show that the DGSM outperforms state-of-the-art baselines by a large margin, and it is capable of generating conformations for a broader range of systems such as proteins and multi-molecular complexes.
****
When in Doubt: Neural Non-Parametric Uncertainty Quantification for Epidemic Forecasting
Harshavardhan Kamarthi, Lingkai Kong, Alexander Rodriguez, Chao Zhang, B. Aditya Prakash
Accurate and trustworthy epidemic forecasting is an important problem for public health planning and disease mitigation. Most existing epidemic forecasting models disregard uncertainty quantification, resulting in mis-calibrated predictions. Recent works in deep neural models for uncertainty-aware time-series forecasting also have several limitations; e.g., it is difficult to specify proper priors in Bayesian NNs, while methods like deep ensembling can be computationally expensive. In this paper, we propose to use neural functional processes to fill this gap. We model epidemic time-series with a probabilistic generative process and propose a functional neural process model called EpiFNP, which directly models the probability distribution of the forecast value in a non-parametric way. In EpiFNP, we use a dynamic stochastic correlation graph to model the correlations between sequences, and design different stochastic latent variables to capture functional uncertainty from different perspectives. Our experiments in a real-time flu forecasting setting show that EpiFNP significantly outperforms state-of-the-art models in both accuracy and calibration metrics, up to 2.5x in accuracy and 2.4x in calibration. Additionally, as EpiFNP learns the relations between the current season and similar patterns of historical seasons, it enables interpretable forecasts. Beyond epidemic forecasting, EpiFNP can be of independent interest for advancing uncertainty quantification in deep sequential models for predictiv

e analytics.
****
Bounds all around: training energy-based models with bidirectional bounds
Cong Geng, Jia Wang, Zhiyong Gao, Jes Frellsen, Søren Hauberg
Energy-based models (EBMs) provide an elegant framework for density estimation, but they are notoriously difficult to train. Recent work has established links to generative adversarial networks, where the EBM is trained through a minimax game with a variational value function. We propose a bidirectional bound on the EBM log-likelihood, such that we maximize a lower bound and minimize an upper bound when solving the minimax game. We link one bound to a gradient penalty that stabilizes training, thereby provide grounding for best engineering practice. To evaluate the bounds we develop a new and efficient estimator of the Jacobi-determinant of the EBM generator. We demonstrate that these developments stabilize training and yield high-quality density estimation and sample generation.
****
CogView: Mastering Text-to-Image Generation via Transformers
Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, Jie Tang
Text-to-Image generation in the general domain has long been an open problem, which requires both a powerful generative model and cross-modal understanding. We propose CogView, a 4-billion-parameter Transformer with VQ-VAE tokenizer to advance this problem. We also demonstrate the finetuning strategies for various downstream tasks, e.g. style learning, super-resolution, text-image ranking and fashion design, and methods to stabilize pretraining, e.g. eliminating NaN losses. CogView achieves the state-of-the-art FID on the blurred MS COCO dataset, outperforming previous GAN-based models and a recent similar work DALL-E.
****
Time-independent Generalization Bounds for SGLD in Non-convex Settings
Tyler Farghly, Patrick Rebeschini
We establish generalization error bounds for stochastic gradient Langevin dynamics (SGLD) with constant learning rate under the assumptions of dissipativity and smoothness, a setting that has received increased attention in the sampling/optimization literature. Unlike existing bounds for SGLD in non-convex settings, ours are time-independent and decay to zero as the sample size increases. Using the framework of uniform stability, we establish time-independent bounds by exploiting the Wasserstein contraction property of the Langevin diffusion, which also allows us to circumvent the need to bound gradients using Lipschitz-like assumptions. Our analysis also supports variants of SGLD that use different discretization methods, incorporate Euclidean projections, or use non-isotropic noise.
****
Nonuniform Negative Sampling and Log Odds Correction with Rare Events Data
HaiYing Wang, Aonan Zhang, Chong Wang
We investigate the issue of parameter estimation with nonuniform negative sampling for imbalanced data. We first prove that, with imbalanced data, the available information about unknown parameters is only tied to the relatively small number of positive instances, which justifies the usage of negative sampling. However, if the negative instances are subsampled to the same level of the positive cases, there is information loss. To maintain more information, we derive the asymptotic distribution of a general inverse probability weighted (IPW) estimator and obtain the optimal sampling probability that minimizes its variance. To further improve the estimation efficiency over the IPW method, we propose a likelihood-based estimator by correcting log odds for the sampled data and prove that the improved estimator has the smallest asymptotic variance among a large class of estimators. It is also more robust to pilot misspecification. We validate our approach on simulated data as well as a real click-through rate dataset with more than 0.3 trillion instances, collected over a period of a month. Both theoretical and empirical results demonstrate the effectiveness of our method.
****
Algorithmic stability and generalization of an unsupervised feature selection algorithm

xinxing wu, Qiang Cheng
Feature selection, as a vital dimension reduction technique, reduces data dimension by identifying an essential subset of input features, which can facilitate interpretable insights into learning and inference processes. Algorithmic stability is a key characteristic of an algorithm regarding its sensitivity to perturbations of input samples. In this paper, we propose an innovative unsupervised feature selection algorithm attaining this stability with provable guarantees. The architecture of our algorithm consists of a feature scorer and a feature selector. The scorer trains a neural network (NN) to globally score all the features, and the selector adopts a dependent sub-NN to locally evaluate the representation abilities for selecting features. Further, we present algorithmic stability analysis and show that our algorithm has a performance guarantee via a generalization error bound. Extensive experimental results on real-world datasets demonstrate superior generalization performance of our proposed algorithm to strong baseline methods. Also, the properties revealed by our theoretical analysis and the stability of our algorithm-selected features are empirically confirmed.
****
On learning sparse vectors from mixture of responses
Nikita Polyanskii
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 and Alignment of Gradient Descent with Random Backpropagation Weights
Ganlin Song, Ruitu Xu, John Lafferty
Stochastic gradient descent with backpropagation is the workhorse of artificial neural networks. It has long been recognized that backpropagation fails to be a biologically plausible algorithm. Fundamentally, it is a non-local procedure---updating one neuron's synaptic weights requires knowledge of synaptic weights or receptive fields of downstream neurons. This limits the use of artificial neural networks as a tool for understanding the biological principles of information processing in the brain. Lillicrap et al. (2016) propose a more biologically plausible "feedback alignment" algorithm that uses random and fixed backpropagation weights, and show promising simulations. In this paper we study the mathematical properties of the feedback alignment procedure by analyzing convergence and alignment for two-layer networks under squared error loss. In the overparameterized setting, we prove that the error converges to zero exponentially fast, and also that regularization is necessary in order for the parameters to become aligned with the random backpropagation weights. Simulations are given that are consistent with this analysis and suggest further generalizations. These results contribute to our understanding of how biologically plausible algorithms might carry out weight learning in a manner different from Hebbian learning, with performance that is comparable with the full non-local backpropagation algorithm.
****
Adder Attention for Vision Transformer
Han Shu, Jiahao Wang, Hanting Chen, Lin Li, Yujiu Yang, Yunhe Wang
Transformer is a new kind of calculation paradigm for deep learning which has shown strong performance on a large variety of computer vision tasks. However, compared with conventional deep models (e.g., convolutional neural networks), vision transformers require more computational resources which cannot be easily deployed on mobile devices. To this end, we present to reduce the energy consumptions using adder neural network (AdderNet). We first theoretically analyze the mechanism of self-attention and the difficulty for applying adder operation into this module. Specifically, the feature diversity, i.e., the rank of attention map using only additions cannot be well preserved. Thus, we develop an adder attention layer that includes an additional identity mapping. With the new operation, vision transformers constructed using additions can also provide powerful feature representations. Experimental results on several benchmarks demonstrate that the

proposed approach can achieve highly competitive performance to that of the baselines while achieving an about 2~3× reduction on the energy consumption.
****
Reverse engineering learned optimizers reveals known and novel mechanisms
Niru Maheswaranathan, David Sussillo, Luke Metz, Ruoxi Sun, Jascha Sohl-Dickstein
Learned optimizers are parametric algorithms that can themselves be trained to solve optimization problems. In contrast to baseline optimizers (such as momentum or Adam) that use simple update rules derived from theoretical principles, learned optimizers use flexible, high-dimensional, nonlinear parameterizations. Although this can lead to better performance, their inner workings remain a mystery. How is a given learned optimizer able to outperform a well tuned baseline? Has it learned a sophisticated combination of existing optimization techniques, or is it implementing completely new behavior? In this work, we address these questions by careful analysis and visualization of learned optimizers. We study learned optimizers trained from scratch on four disparate tasks, and discover that they have learned interpretable behavior, including: momentum, gradient clipping, learning rate schedules, and new forms of learning rate adaptation. Moreover, we show how dynamics and mechanisms inside of learned optimizers orchestrate these computations. Our results help elucidate the previously murky understanding of how learned optimizers work, and establish tools for interpreting future learned optimizers.
****
Matching a Desired Causal State via Shift Interventions
Jiaqi Zhang, Chandler Squires, Caroline Uhler
Transforming a causal system from a given initial state to a desired target state is an important task permeating multiple fields including control theory, biology, and materials science. In causal models, such transformations can be achieved by performing a set of interventions. In this paper, we consider the problem of identifying a shift intervention that matches the desired mean of a system through active learning. We define the Markov equivalence class that is identifiable from shift interventions and propose two active learning strategies that are guaranteed to exactly match a desired mean. We then derive a worst-case lower bound for the number of interventions required and show that these strategies are optimal for certain classes of graphs. In particular, we show that our strategies may require exponentially fewer interventions than the previously considered approaches, which optimize for structure learning in the underlying causal graph. In line with our theoretical results, we also demonstrate experimentally that our proposed active learning strategies require fewer interventions compared to several baselines.
****
Unsupervised Noise Adaptive Speech Enhancement by Discriminator-Constrained Optimal Transport
Hsin-Yi Lin, Huan-Hsin Tseng, Xugang Lu, Yu Tsao
This paper presents a novel discriminator-constrained optimal transport network (DOTN) that performs unsupervised domain adaptation for speech enhancement (SE), which is an essential regression task in speech processing. The DOTN aims to estimate clean references of noisy speech in a target domain, by exploiting the knowledge available from the source domain. The domain shift between training and testing data has been reported to be an obstacle to learning problems in diverse fields. Although rich literature exists on unsupervised domain adaptation for classification, the methods proposed, especially in regressions, remain scarce and often depend on additional information regarding the input data. The proposed DOTN approach tactically fuses the optimal transport (OT) theory from mathematical analysis with generative adversarial frameworks, to help evaluate continuous labels in the target domain. The experimental results on two SE tasks demonstrate that by extending the classical OT formulation, our proposed DOTN outperforms previous adversarial domain adaptation frameworks in a purely unsupervised manner.
****

Optimality of variational inference for stochasticblock model with missing links
Solenne Gaucher, Olga Klopp
Variational methods are extremely popular in the analysis of network data. Statistical guarantees obtained for these methods typically provide asymptotic normality for the problem of estimation of global model parameters under the stochastic block model. In the present work, we consider the case of networks with missing links that is important in application and show that the variational approximation to the maximum likelihood estimator converges at the minimax rate. This provides the first minimax optimal and tractable estimator for the problem of parameter estimation for the stochastic block model with missing links. We complement our results with numerical studies of simulated and real networks, which confirm the advantages of this estimator over current methods.
****
Policy Learning Using Weak Supervision
Jingkang Wang, Hongyi Guo, Zhaowei Zhu, Yang Liu
Most existing policy learning solutions require the learning agents to receive high-quality supervision signals, e.g., rewards in reinforcement learning (RL) or high-quality expert demonstrations in behavioral cloning (BC). These quality supervisions are either infeasible or prohibitively expensive to obtain in practice. We aim for a unified framework that leverages the available cheap weak supervisions to perform policy learning efficiently. To handle this problem, we treat the weak supervision'' as imperfect information coming from a peer agent, and evaluate the learning agent's policy based on a correlated agreement'' with the peer agent's policy (instead of simple agreements). Our approach explicitly punishes a policy for overfitting to the weak supervision. In addition to theoretical guarantees, extensive evaluations on tasks including RL with noisy reward, BC with weak demonstrations, and standard policy co-training (RL + BC) show that our method leads to substantial performance improvements, especially when the complexity or the noise of the learning environments is high.
****
Chasing Sparsity in Vision Transformers: An End-to-End Exploration
Tianlong Chen, Yu Cheng, Zhe Gan, Lu Yuan, Lei Zhang, Zhangyang 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.
****
Graphical Models in Heavy-Tailed Markets
Jose Vinicius de Miranda Cardoso, Jiaxi Ying, 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-authors prior to requesting a name change in the electronic proceedings.
****
A Shading-Guided Generative Implicit Model for Shape-Accurate 3D-Aware Image Synthesis
Xingang Pan, Xudong XU, Chen Change Loy, Christian Theobalt, Bo Dai
The advancement of generative radiance fields has pushed the boundary of 3D-aware image synthesis. Motivated by the observation that a 3D object should look realistic from multiple viewpoints, these methods introduce a multi-view constraint as regularization to learn valid 3D radiance fields from 2D images. Despite the progress, they often fall short of capturing accurate 3D shapes due to the shape-color ambiguity, limiting their applicability in downstream tasks. In this work, we address this ambiguity by proposing a novel shading-guided generative implicit model that is able to learn a starkly improved shape representation. Our key insight is that an accurate 3D shape should also yield a realistic rendering under different lighting conditions. This multi-lighting constraint is realized by modeling illumination explicitly and performing shading with various lighting conditions. Gradients are derived by feeding the synthesized images to a discriminator. To compensate for the additional computational burden of calculating sur

face normals, we further devise an efficient volume rendering strategy via surface tracking, reducing the training and inference time by 24% and 48%, respectively. Our experiments on multiple datasets show that the proposed approach achieves photorealistic 3D-aware image synthesis while capturing accurate underlying 3D shapes. We demonstrate improved performance of our approach on 3D shape reconstruction against existing methods, and show its applicability on image relighting. Our code is available at https://github.com/XingangPan/ShadeGAN.
****

XCiT: Cross-Covariance Image Transformers

Alaaeldin Ali, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek, Herve Jegou

Following their success in natural language processing, transformers have recently shown much promise for computer vision. The self-attention operation underlying transformers yields global interactions between all tokens ,i.e. words or image patches, and enables flexible modelling of image data beyond the local interactions of convolutions. This flexibility, however, comes with a quadratic complexity in time and memory, hindering application to long sequences and high-resolution images. We propose a “transposed” version of self-attention that operates across feature channels rather than tokens, where the interactions are based on the cross-covariance matrix between keys and queries. The resulting cross-covariance attention (XCA) has linear complexity in the number of tokens, and allows efficient processing of high-resolution images.Our cross-covariance image transformer (XCiT) is built upon XCA. It combines the accuracy of conventional transformers with the scalability of convolutional architectures. We validate the effectiveness and generality of XCiT by reporting excellent results on multiple vision benchmarks, including image classification and self-supervised feature learning on ImageNet-1k, object detection and instance segmentation on COCO, and semantic segmentation on ADE20k.We will opensource our code and trained models to reproduce the reported results.
****

Row-clustering of a Point Process-valued Matrix

Lihao Yin, Ganggang Xu, Huiyan Sang, Yongtao Guan

Structured point process data harvested from various platforms poses new challenges to the machine learning community. To cluster repeatedly observed marked point processes, we propose a novel mixture model of multi-level marked point processes for identifying potential heterogeneity in the observed data. Specifically, we study a matrix whose entries are marked log-Gaussian Cox processes and cluster rows of such a matrix. An efficient semi-parametric Expectation-Solution (ES) algorithm combined with functional principal component analysis (FPCA) of point processes is proposed for model estimation. The effectiveness of the proposed framework is demonstrated through simulation studies and real data analyses.
****

Fine-Grained Neural Network Explanation by Identifying Input Features with Predictive Information

Yang Zhang, Ashkan Khakzar, Yawei Li, Azade Farshad, Seong Tae Kim, Nassir Navab

One principal approach for illuminating a black-box neural network is feature attribution, i.e. identifying the importance of input features for the network’s prediction. The predictive information of features is recently proposed as a proxy for the measure of their importance. So far, the predictive information is only identified for latent features by placing an information bottleneck within the network. We propose a method to identify features with predictive information in the input domain. The method results in fine-grained identification of input features' information and is agnostic to network architecture. The core idea of our method is leveraging a bottleneck on the input that only lets input features associated with predictive latent features pass through. We compare our method with several feature attribution methods using mainstream feature attribution evaluation experiments. The code is publicly available.
****

Fast Minimum-norm Adversarial Attacks through Adaptive Norm Constraints

Maura Pintor, Fabio Roli, Wieland Brendel, Battista Biggio
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Uncertainty Quantification and Deep Ensembles
Rahul Rahaman, alexandre thiery
Deep Learning methods are known to suffer from calibration issues: they typically produce over-confident estimates. These problems are exacerbated in the low data regime. Although the calibration of probabilistic models is well studied, calibrating extremely over-parametrized models in the low-data regime presents unique challenges. We show that deep-ensembles do not necessarily lead to improved calibration properties. In fact, we show that standard ensembling methods, when used in conjunction with modern techniques such as mixup regularization, can lead to less calibrated models. This text examines the interplay between three of the most simple and commonly used approaches to leverage deep learning when data is scarce: data-augmentation, ensembling, and post-processing calibration methods. We demonstrate that, although standard ensembling techniques certainly help to boost accuracy, the calibration of deep ensembles relies on subtle trade-offs. We also find that calibration methods such as temperature scaling need to be slightly tweaked when used with deep-ensembles and, crucially, need to be executed after the averaging process. Our simulations indicate that, in the low data regime, this simple strategy can halve the Expected Calibration Error (ECE) on a range of benchmark classification problems when compared to standard deep-ensembles.
****
Directed Probabilistic Watershed
Enrique Fita Sanmartin, Sebastian Damrich, Fred A. Hamprecht
The Probabilistic Watershed is a semi-supervised learning algorithm applied on undirected graphs. Given a set of labeled nodes (seeds), it defines a Gibbs probability distribution over all possible spanning forests disconnecting the seeds. It calculates, for every node, the probability of sampling a forest connecting a certain seed with the considered node. We propose the "Directed Probabilistic Watershed", an extension of the Probabilistic Watershed algorithm to directed graphs. Building on the Probabilistic Watershed, we apply the Matrix Tree Theorem for directed graphs and define a Gibbs probability distribution over all incoming directed forests rooted at the seeds. Similar to the undirected case, this turns out to be equivalent to the Directed Random Walker. Furthermore, we show that in the limit case in which the Gibbs distribution has infinitely low temperature, the labeling of the Directed Probabilistic Watershed is equal to the one induced by the incoming directed forest of minimum cost. Finally, for illustration, we compare the empirical performance of the proposed method with other semi-supervised segmentation methods for directed graphs.
****
Laplace Redux - Effortless Bayesian Deep Learning
Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, Philipp Hennig
Bayesian formulations of deep learning have been shown to have compelling theoretical properties and offer practical functional benefits, such as improved predictive uncertainty quantification and model selection. The Laplace approximation (LA) is a classic, and arguably the simplest family of approximations for the intractable posteriors of deep neural networks. Yet, despite its simplicity, the LA is not as popular as alternatives like variational Bayes or deep ensembles. This may be due to assumptions that the LA is expensive due to the involved Hessian computation, that it is difficult to implement, or that it yields inferior results. In this work we show that these are misconceptions: we (i) review the range of variants of the LA including versions with minimal cost overhead; (ii) introduce "laplace", an easy-to-use software library for PyTorch offering user-friendly access to all major flavors of the LA; and (iii) demonstrate through extensi

ve experiments that the LA is competitive with more popular alternatives in terms of performance, while excelling in terms of computational cost. We hope that this work will serve as a catalyst to a wider adoption of the LA in practical deep learning, including in domains where Bayesian approaches are not typically considered at the moment.
****
Hessian Eigenspectra of More Realistic Nonlinear Models
Zhenyu Liao, Michael W. Mahoney
Given an optimization problem, the Hessian matrix and its eigenspectrum can be used in many ways, ranging from designing more efficient second-order algorithms to performing model analysis and regression diagnostics. When nonlinear models and non-convex problems are considered, strong simplifying assumptions are often made to make Hessian spectral analysis more tractable.This leads to the question of how relevant the conclusions of such analyses are for realistic nonlinear models. In this paper, we exploit tools from random matrix theory to make a precise characterization of the Hessian eigenspectra for a broad family of nonlinear models that extends the classical generalized linear models, without relying on strong simplifying assumptions used previously. We show that, depending on the data properties, the nonlinear response model, and the loss function, the Hessian can have qualitatively different spectral behaviors: of bounded or unbounded support, with single- or multi-bulk, and with isolated eigenvalues on the left- or right-hand side of the main eigenvalue bulk. By focusing on such a simple but nontrivial model, our analysis takes a step forward to unveil the theoretical origin of many visually striking features observed in more realistic machine learning models.
****
Explicable Reward Design for Reinforcement Learning Agents
Rati Devidze, Goran Radanovic, Parameswaran Kamalaruban, Adish Singla
We study the design of explicable reward functions for a reinforcement learning agent while guaranteeing that an optimal policy induced by the function belongs to a set of target policies. By being explicable, we seek to capture two properties: (a) informativeness so that the rewards speed up the agent's convergence, and (b) sparseness as a proxy for ease of interpretability of the rewards. The key challenge is that higher informativeness typically requires dense rewards for many learning tasks, and existing techniques do not allow one to balance these two properties appropriately. In this paper, we investigate the problem from the perspective of discrete optimization and introduce a novel framework, ExpRD, to design explicable reward functions. ExpRD builds upon an informativeness criterion that captures the (sub-)optimality of target policies at different time horizons in terms of actions taken from any given starting state. We provide a mathematical analysis of ExpRD, and show its connections to existing reward design techniques, including potential-based reward shaping. Experimental results on two navigation tasks demonstrate the effectiveness of ExpRD in designing explicable reward functions.
****
A Minimalist Approach to Offline Reinforcement Learning
Scott Fujimoto, Shixiang (Shane) Gu
Offline reinforcement learning (RL) defines the task of learning from a fixed batch of data. Due to errors in value estimation from out-of-distribution actions, most offline RL algorithms take the approach of constraining or regularizing the policy with the actions contained in the dataset. Built on pre-existing RL algorithms, modifications to make an RL algorithm work offline comes at the cost of additional complexity. Offline RL algorithms introduce new hyperparameters and often leverage secondary components such as generative models, while adjusting the underlying RL algorithm. In this paper we aim to make a deep RL algorithm work while making minimal changes. We find that we can match the performance of state-of-the-art offline RL algorithms by simply adding a behavior cloning term to the policy update of an online RL algorithm and normalizing the data. The resulting algorithm is a simple to implement and tune baseline, while more than halving the overall run time by removing the additional computational overheads of pre

vious methods.
****
SIMONe: View-Invariant, Temporally-Abstracted Object Representations via Unsupervised Video Decomposition
Rishabh Kabra, Daniel Zoran, Goker Erdogan, Loic Matthey, Antonia Creswell, Matt Botvinick, Alexander Lerchner, Chris Burgess
To help agents reason about scenes in terms of their building blocks, we wish to extract the compositional structure of any given scene (in particular, the configuration and characteristics of objects comprising the scene). This problem is especially difficult when scene structure needs to be inferred while also estimating the agent’s location/viewpoint, as the two variables jointly give rise to the agent’s observations. We present an unsupervised variational approach to this problem. Leveraging the shared structure that exists across different scenes, our model learns to infer two sets of latent representations from RGB video input alone: a set of "object" latents, corresponding to the time-invariant, object-level contents of the scene, as well as a set of "frame" latents, corresponding to global time-varying elements such as viewpoint. This factorization of latents allows our model, SIMONe, to represent object attributes in an allocentric manner which does not depend on viewpoint. Moreover, it allows us to disentangle object dynamics and summarize their trajectories as time-abstracted, view-invariant, per-object properties. We demonstrate these capabilities, as well as the model's performance in terms of view synthesis and instance segmentation, across three procedurally generated video datasets.
****
Simple Stochastic and Online Gradient Descent Algorithms for Pairwise Learning
ZHENHUAN YANG, Yunwen Lei, Puyu Wang, Tianbao Yang, Yiming Ying
Pairwise learning refers to learning tasks where the loss function depends on a pair of instances. It instantiates many important machine learning tasks such as bipartite ranking and metric learning. A popular approach to handle streaming data in pairwise learning is an online gradient descent (OGD) algorithm, where one needs to pair the current instance with a buffering set of previous instances with a sufficiently large size and therefore suffers from a scalability issue. In this paper, we propose simple stochastic and online gradient descent methods for pairwise learning. A notable difference from the existing studies is that we only pair the current instance with the previous one in building a gradient direction, which is efficient in both the storage and computational complexity. We develop novel stability results, optimization, and generalization error bounds for both convex and nonconvex as well as both smooth and nonsmooth problems. We introduce novel techniques to decouple the dependency of models and the previous instance in both the optimization and generalization analysis. Our study resolves an open question on developing meaningful generalization bounds for OGD using a buffering set with a very small fixed size. We also extend our algorithms and stability analysis to develop differentially private SGD algorithms for pairwise learning which significantly improves the existing results.
****
User-Level Differentially Private Learning via Correlated Sampling
Badih Ghazi, Ravi Kumar, Pasin Manurangsi
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Asynchronous Decentralized Online Learning
Jiyan Jiang, Wenpeng Zhang, Jinjie GU, Wenwu Zhu
Most existing algorithms in decentralized online learning are conducted in the synchronous setting. However, synchronization makes these algorithms suffer from the straggler problem, i.e., fast learners have to wait for slow learners, which significantly reduces such algorithms' overall efficiency. To overcome this problem, we study decentralized online learning in the asynchronous setting, which allows different learners to work at their own pace. We first formulate the fram

ework of Asynchronous Decentralized Online Convex Optimization, which specifies the whole process of asynchronous decentralized online learning using a sophisticated event indexing system. Then we propose the Asynchronous Decentralized Online Gradient-Push (AD-OGP) algorithm, which performs asymmetric gossiping communication and instantaneous model averaging. We further derive a regret bound of AD-OGP, which is a function of the network topology, the levels of processing delays, and the levels of communication delays. Extensive experiments show that AD-OGP runs significantly faster than its synchronous counterpart and also verify the theoretical results.
****
Multi-Step Budgeted Bayesian Optimization with Unknown Evaluation Costs
Raul Astudillo, Daniel Jiang, Maximilian Balandat, Eytan Bakshy, Peter Frazier
Bayesian optimization (BO) is a sample-efficient approach to optimizing costly-to-evaluate black-box functions. Most BO methods ignore how evaluation costs may vary over the optimization domain. However, these costs can be highly heterogeneous and are often unknown in advance in many practical settings, such as hyperparameter tuning of machine learning algorithms or physics-based simulation optimization. Moreover, those few existing methods that acknowledge cost heterogeneity do not naturally accommodate a budget constraint on the total evaluation cost. This combination of unknown costs and a budget constraint introduces a new dimension to the exploration-exploitation trade-off, where learning about the cost incurs a cost itself. Existing methods do not reason about the various trade-offs of this problem in a principled way, leading often to poor performance. We formalize this claim by proving that the expected improvement and the expected improvement per unit of cost, arguably the two most widely used acquisition functions in practice, can be arbitrarily inferior with respect to the optimal non-myopic policy. To overcome the shortcomings of existing approaches, we propose the budgeted multi-step expected improvement, a non-myopic acquisition function that generalizes classical expected improvement to the setting of heterogeneous and unknown evaluation costs. We show that our acquisition function outperforms existing methods in a variety of synthetic and real problems.
****
Model-Based Domain Generalization
Alexander Robey, George J. Pappas, Hamed Hassani
Despite remarkable success in a variety of applications, it is well-known that deep learning can fail catastrophically when presented with out-of-distribution data. Toward addressing this challenge, we consider the \emph{domain generalization} problem, wherein predictors are trained using data drawn from a family of related training domains and then evaluated on a distinct and unseen test domain. We show that under a natural model of data generation and a concomitant invariance condition, the domain generalization problem is equivalent to an infinite-dimensional constrained statistical learning problem; this problem forms the basis of our approach, which we call Model-Based Domain Generalization. Due to the inherent challenges in solving constrained optimization problems in deep learning, we exploit nonconvex duality theory to develop unconstrained relaxations of this statistical problem with tight bounds on the duality gap. Based on this theoretical motivation, we propose a novel domain generalization algorithm with convergence guarantees.   In our experiments, we report improvements of up to 30% over state-of-the-art domain generalization baselines on several benchmarks including ColoredMNIST, Camelyon17-WILDS, FMoW-WILDS, and PACS.
****
α-IoU: A Family of Power Intersection over Union Losses for Bounding Box Regression
JIABO HE, Sarah Erfani, Xingjun Ma, James Bailey, Ying Chi, Xian-Sheng Hua
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Practical Large-Scale Linear Programming using Primal-Dual Hybrid Gradient

David Applegate, Mateo Diaz, Oliver Hinder, Haihao Lu, Miles Lubin, Brendan O'Do
noghue, Warren Schudy
Requests for name changes in the electronic proceedings will be accepted with 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 the Provable Generalization of Recurrent Neural Networks
Lifu Wang, Bo Shen, Bo Hu, Xing Cao
Requests for name changes in the electronic proceedings will be accepted with 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 Spline Approximations
Minsu Cho, Aditya Balu, Ameya Joshi, Anjana Deva Prasad, Biswajit Khara, Soumik
Sarkar, Baskar Ganapathysubramanian, Adarsh Krishnamurthy, Chinmay Hegde
The paradigm of differentiable programming has significantly enhanced the scope
of machine learning via the judicious use of gradient-based optimization. Howeve
r, standard differentiable programming methods (such as autodiff) typically requ
ire that the machine learning models be differentiable, limiting their applicabi
lity. Our goal in this paper is to use a new, principled approach to extend grad
ient-based optimization to functions well modeled by splines, which encompass a
large family of piecewise polynomial models. We derive the form of the (weak) Ja
cobian of such functions and show that it exhibits a block-sparse structure that
 can be computed implicitly and efficiently. Overall, we show that leveraging th
is redesigned Jacobian in the form of a differentiable "layer'' in predictive mo
dels leads to improved performance in diverse applications such as image segment
ation, 3D point cloud reconstruction, and finite element analysis. We also open-
source the code at \url{https://github.com/idealab-isu/DSA}.
****
Rate-Optimal Subspace Estimation on Random Graphs
Zhixin Zhou, Fan Zhou, Ping Li, Cun-Hui 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.
****
Estimating the Unique Information of Continuous Variables
Ari Pakman, Amin Nejatbakhsh, Dar Gilboa, Abdullah Makkeh, Luca Mazzucato, Micha
el Wibral, Elad Schneidman
The integration and transfer of information from multiple sources to multiple ta
rgets is a core motive of neural systems. The emerging field of partial informat
ion decomposition (PID) provides a novel information-theoretic lens into these m
echanisms by identifying synergistic, redundant, and unique contributions to the
 mutual information between one and several variables. While many works have stu
died aspects of PID for Gaussian and discrete distributions, the case of general
 continuous distributions is still uncharted territory. In this work we present
a method for estimating the unique information in continuous distributions, for
the case of one versus two variables. Our method solves the associated optimizat
ion problem over the space of distributions with fixed bivariate marginals by c
ombining copula decompositions and techniques developed to optimize variational
autoencoders. We obtain excellent agreement with known analytic results for Gaus
sians, and illustrate the power of our new approach in several brain-inspired n
eural models. Our method is capable of recovering the effective connectivity of
a chaotic network of rate neurons, and uncovers a complex trade-off between redu
ndancy, synergy and unique information in recurrent networks trained to solve a
generalized XOR~task.
****
Reliable Causal Discovery with Improved Exact Search and Weaker Assumptions

Ignavier Ng, Yujia Zheng, Jiji Zhang, Kun Zhang
Many of the causal discovery methods rely on the faithfulness assumption to guarantee asymptotic correctness. However, the assumption can be approximately violated in many ways, leading to sub-optimal solutions. Although there is a line of research in Bayesian network structure learning that focuses on weakening the assumption, such as exact search methods with well-defined score functions, they do not scale well to large graphs. In this work, we introduce several strategies to improve the scalability of exact score-based methods in the linear Gaussian setting. In particular, we develop a super-structure estimation method based on the support of inverse covariance matrix which requires assumptions that are strictly weaker than faithfulness, and apply it to restrict the search space of exact search. We also propose a local search strategy that performs exact search on the local clusters formed by each variable and its neighbors within two hops in the super-structure. Numerical experiments validate the efficacy of the proposed procedure, and demonstrate that it scales up to hundreds of nodes with a high accuracy.
****
Node Dependent Local Smoothing for Scalable Graph Learning
Wentao Zhang, Mingyu Yang, Zeang Sheng, Yang Li, Wen Ouyang, Yangyu Tao, Zhi Yang, Bin CUI
Recent works reveal that feature or label smoothing lies at the core of Graph Neural Networks (GNNs). Concretely, they show feature smoothing combined with simple linear regression achieves comparable performance with the carefully designed GNNs, and a simple MLP model with label smoothing of its prediction can outperform the vanilla GCN. Though an interesting finding, smoothing has not been well understood, especially regarding how to control the extent of smoothness. Intuitively, too small or too large smoothing iterations may cause under-smoothing or over-smoothing and can lead to sub-optimal performance. Moreover, the extent of smoothness is node-specific, depending on its degree and local structure. To this end, we propose a novel algorithm called node-dependent local smoothing (NDLS), which aims to control the smoothness of every node by setting a node-specific smoothing iteration. Specifically, NDLS computes influence scores based on the adjacency matrix and selects the iteration number by setting a threshold on the scores. Once selected, the iteration number can be applied to both feature smoothing and label smoothing. Experimental results demonstrate that NDLS enjoys high accuracy -- state-of-the-art performance on node classifications tasks, flexibility -- can be incorporated with any models, scalability and efficiency -- can support large scale graphs with fast training.
****
Parallel and Efficient Hierarchical k-Median Clustering
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-authors prior to requesting a name change in the electronic proceedings.
****
Human-Adversarial Visual Question Answering
Sasha Sheng, Amanpreet Singh, Vedanuj Goswami, Jose Magana, Tristan Thrush, Wojciech Galuba, Devi Parikh, Douwe Kiela
Performance on the most commonly used Visual Question Answering dataset (VQA v2) is starting to approach human accuracy. However, in interacting with state-of-the-art VQA models, it is clear that the problem is far from being solved. In order to stress test VQA models, we benchmark them against human-adversarial examples. Human subjects interact with a state-of-the-art VQA model, and for each image in the dataset, attempt to find a question where the model’s predicted answer is incorrect. We find that a wide range of state-of-the-art models perform poorly when evaluated on these examples. We conduct an extensive analysis of the collected adversarial examples and provide guidance on future research directions. We hope that this Adversarial VQA (AdVQA) benchmark can help drive progress in th

e field and advance the state of the art.
****
Across-animal odor decoding by probabilistic manifold alignment
Pedro Herrero-Vidal, Dmitry Rinberg, Cristina Savin
Identifying the common structure of neural dynamics across subjects is key for e xtracting unifying principles of brain computation and for many brain machine in terface applications. Here, we propose a novel probabilistic approach for aligni ng stimulus-evoked responses from multiple animals in a common low dimensional m anifold and use hierarchical inference to identify which stimulus drives neural activity in any given trial. Our probabilistic decoder is robust to a range of f eatures of the neural responses and significantly outperforms existing neural al ignment procedures. When applied to recordings from the mouse olfactory bulb, ou r approach reveals low-dimensional population dynamics that are odor specific an d have consistent structure across animals. Thus, our decoder can be used for in creasing the robustness and scalability of neural-based chemical detection.
****
Excess Capacity and Backdoor Poisoning
Naren Manoj, Avrim Blum
A backdoor data poisoning attack is an adversarial attack wherein the attacker i njects several watermarked, mislabeled training examples into a training set. Th e watermark does not impact the test-time performance of the model on typical da ta; however, the model reliably errs on watermarked examples.To gain a better fo undational understanding of backdoor data poisoning attacks, we present a formal theoretical framework within which one can discuss backdoor data poisoning atta cks for classification problems. We then use this to analyze important statistic al and computational issues surrounding these attacks.On the statistical front, we identify a parameter we call the memorization capacity that captures the intr insic vulnerability of a learning problem to a backdoor attack. This allows us t o argue about the robustness of several natural learning problems to backdoor at tacks. Our results favoring the attacker involve presenting explicit constructio ns of backdoor attacks, and our robustness results show that some natural proble m settings cannot yield successful backdoor attacks.From a computational standpo int, we show that under certain assumptions, adversarial training can detect the presence of backdoors in a training set. We then show that under similar assump tions, two closely related problems we call backdoor filtering and robust genera lization are nearly equivalent. This implies that it is both asymptotically nece ssary and sufficient to design algorithms that can identify watermarked examples in the training set in order to obtain a learning algorithm that both generaliz es well to unseen data and is robust to backdoors.
****
A Convergence Analysis of Gradient Descent on Graph Neural Networks
Pranjal Awasthi, Abhimanyu Das, Sreenivas Gollapudi
Requests for name changes in the electronic proceedings will be accepted with 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 rendering with perturbed optimizers
Quentin Le Lidec, Ivan Laptev, Cordelia Schmid, Justin Carpentier
Reasoning about 3D scenes from their 2D image projections is one of the core pro blems in computer vision. Solutions to this inverse and ill-posed problem typica lly involve a search for models that best explain observed image data. Notably, images depend both on the properties of observed scenes and on the process of im age formation. Hence, if optimization techniques should be used to explain image s, it is crucial to design differentable functions for the projection of 3D scen es into images, also known as differentiable rendering. Previous approaches to d ifferentiable rendering typically replace non-differentiable operations by smoot h approximations, impacting the subsequent 3D estimation. In this paper, we take a more general approach and study differentiable renderers through the prism of randomized optimization and the related notion of perturbed optimizers. In part

icular, our work highlights the link between some well-known differentiable renderer formulations and randomly smoothed optimizers, and introduces differentiable perturbed renderers. We also propose a variance reduction mechanism to alleviate the computational burden inherent to perturbed optimizers and introduce an adaptive scheme to automatically adjust the smoothing parameters of the rendering process. We apply our method to 3D scene reconstruction and demonstrate its advantages on the tasks of 6D pose estimation and 3D mesh reconstruction. By providing informative gradients that can be used as a strong supervisory signal, we demonstrate the benefits of perturbed renderers to obtain more accurate solutions when compared to the state-of-the-art alternatives using smooth gradient approximations.
****
BCORLE(λ): An Offline Reinforcement Learning and Evaluation Framework for Coupons Allocation in E-commerce Market
Yang Zhang, Bo Tang, Qingyu Yang, Dou An, Hongyin Tang, Chenyang Xi, Xueying LI, Feiyu Xiong
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Nested Variational Inference
Heiko Zimmermann, Hao Wu, Babak Esmaeili, Jan-Willem van de Meent
We develop nested variational inference (NVI), a family of methods that learn proposals for nested importance samplers by minimizing an forward or reverse KL divergence at each level of nesting. NVI is applicable to many commonly-used importance sampling strategies and provides a mechanism for learning intermediate densities, which can serve as heuristics to guide the sampler. Our experiments apply NVI to (a) sample from a multimodal distribution using a learned annealing path (b) learn heuristics that approximate the likelihood of future observations in a hidden Markov model and (c) to perform amortized inference in hierarchical deep generative models. We observe that optimizing nested objectives leads to improved sample quality in terms of log average weight and effective sample size.
****
Exponential Bellman Equation and Improved Regret Bounds for Risk-Sensitive Reinforcement Learning
Yingjie Fei, Zhuoran Yang, Yudong Chen, Zhaoran Wang
We study risk-sensitive reinforcement learning (RL) based on the entropic risk measure. Although existing works have established non-asymptotic regret guarantees for this problem, they leave open an exponential gap between the upper and lower bounds. We identify the deficiencies in existing algorithms and their analysis that result in such a gap. To remedy these deficiencies, we investigate a simple transformation of the risk-sensitive Bellman equations, which we call the exponential Bellman equation. The exponential Bellman equation inspires us to develop a novel analysis of Bellman backup procedures in risk-sensitive RL algorithms, and further motivates the design of a novel exploration mechanism. We show that these analytic and algorithmic innovations together lead to improved regret upper bounds over existing ones.
****
On sensitivity of meta-learning to support data
Mayank Agarwal, Mikhail Yurochkin, Yuekai Sun
Meta-learning algorithms are widely used for few-shot learning. For example, image recognition systems that readily adapt to unseen classes after seeing only a few labeled examples. Despite their success, we show that modern meta-learning algorithms are extremely sensitive to the data used for adaptation, i.e. support data. In particular, we demonstrate the existence of (unaltered, in-distribution, natural) images that, when used for adaptation, yield accuracy as low as 4\% or as high as 95\% on standard few-shot image classification benchmarks. We explain our empirical findings in terms of class margins, which in turn suggests that robust and safe meta-learning requires larger margins than supervised learning.

****
On Large-Cohort Training for Federated Learning
Zachary Charles, Zachary Garrett, Zhouyuan Huo, Sergei Shmulyian, Virginia Smith
Federated learning methods typically learn a model by iteratively sampling updat es from a population of clients. In this work, we explore how the number of clie nts sampled at each round (the cohort size) impacts the quality of the learned m odel and the training dynamics of federated learning algorithms. Our work poses three fundamental questions. First, what challenges arise when trying to scale f ederated learning to larger cohorts? Second, what parallels exist between cohort sizes in federated learning, and batch sizes in centralized learning? Last, how can we design federated learning methods that effectively utilize larger cohort sizes? We give partial answers to these questions based on extensive empirical evaluation. Our work highlights a number of challenges stemming from the use of larger cohorts. While some of these (such as generalization issues and diminishi ng returns) are analogs of large-batch training challenges, others (including ca tastrophic training failures and fairness concerns) are unique to federated lear ning.
****
Generic Neural Architecture Search via Regression
Yuhong Li, Cong Hao, Pan Li, Jinjun Xiong, Deming Chen
Most existing neural architecture search (NAS) algorithms are dedicated to and e valuated by the downstream tasks, e.g., image classification in computer vision. However, extensive experiments have shown that, prominent neural architectures, such as ResNet in computer vision and LSTM in natural language processing, are generally good at extracting patterns from the input data and perform well on di fferent downstream tasks. In this paper, we attempt to answer two fundamental qu estions related to NAS. (1) Is it necessary to use the performance of specific d ownstream tasks to evaluate and search for good neural architectures? (2) Can we perform NAS effectively and efficiently while being agnostic to the downstream tasks? To answer these questions, we propose a novel and generic NAS framework, termed Generic NAS (GenNAS). GenNAS does not use task-specific labels but instea d adopts regression on a set of manually designed synthetic signal bases for arc hitecture evaluation. Such a self-supervised regression task can effectively eva luate the intrinsic power of an architecture to capture and transform the input signal patterns, and allow more sufficient usage of training samples. Extensive experiments across 13 CNN search spaces and one NLP space demonstrate the remark able efficiency of GenNAS using regression, in terms of both evaluating the neur al architectures (quantified by the ranking correlation Spearman's rho between t he approximated performances and the downstream task performances) and the conve rgence speed for training (within a few seconds). For example, on NAS-Bench-101, GenNAS achieves 0.85 rho while the existing efficient methods only achieve 0.38 . We then propose an automatic task search to optimize the combination of synthe tic signals using limited downstream-task-specific labels, further improving the performance of GenNAS. We also thoroughly evaluate GenNAS's generality and end- to-end NAS performance on all search spaces, which outperforms almost all existi ng works with significant speedup. For example, on NASBench-201, GenNAS can find near-optimal architectures within 0.3 GPU hour.
****
The best of both worlds: stochastic and adversarial episodic MDPs with unknown t ransition
Tiancheng Jin, Longbo Huang, Haipeng Luo
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-auth ors prior to requesting a name change in the electronic proceedings.
****
Private learning implies quantum stability
Yihui Quek, Srinivasan Arunachalam, John A Smolin
Learning an unknown n-qubit quantum state rho is a fundamental challenge in quan tum computing. Information-theoretically, it is known that tomography requires e

xponential in n many copies of rho to estimate its entries. Motivated by learning theory, Aaronson et al. introduced many (weaker) learning models: the PAC model of learning states (Proceedings of Royal Society A'07), shadow tomography (STOC'18) for learning shadows" of a state, a model that also requires learners to be differentially private (STOC'19) and the online model of learning states (NeurIPS'18). In these models it was shown that an unknown state can be learnedapproximately" using linear in n many copies of rho. But is there any relationship between these models? In this paper we prove a sequence of (information-theoretic) implications from differentially-private PAC learning to online learning and then to quantum stability.Our main result generalizes the recent work of Bun, Livni and Moran (Journal of the ACM'21) who showed that finite Littlestone dimension (of Boolean-valued concept classes) implies PAC learnability in the (approximate) differentially private (DP) setting. We first consider their work in the real-valued setting and further extend to their techniques to the setting of learning quantum states. Key to our results is our generic quantum online learner, Robust Standard Optimal Algorithm (RSOA), which is robust to adversarial imprecision. We then show information-theoretic implications between DP learning quantum states in the PAC model, learnability of quantum states in the one-way communication model, online learning of quantum states, quantum stability (which is our conceptual contribution), various combinatorial parameters and give further applications to gentle shadow tomography and noisy quantum state learning.
****
Interesting Object, Curious Agent: Learning Task-Agnostic Exploration
Simone Parisi, Victoria Dean, Deepak Pathak, Abhinav Gupta
Common approaches for task-agnostic exploration learn tabula-rasa --the agent assumes isolated environments and no prior knowledge or experience. However, in the real world, agents learn in many environments and always come with prior experiences as they explore new ones. Exploration is a lifelong process. In this paper, we propose a paradigm change in the formulation and evaluation of task-agnostic exploration. In this setup, the agent first learns to explore across many environments without any extrinsic goal in a task-agnostic manner.Later on, the agent effectively transfers the learned exploration policy to better explore new environments when solving tasks. In this context, we evaluate several baseline exploration strategies and present a simple yet effective approach to learning task-agnostic exploration policies. Our key idea is that there are two components of exploration: (1) an agent-centric component encouraging exploration of unseen parts of the environment based on an agent’s belief; (2) an environment-centric component encouraging exploration of inherently interesting objects. We show that our formulation is effective and provides the most consistent exploration across several training-testing environment pairs. We also introduce benchmarks and metrics for evaluating task-agnostic exploration strategies. The source code is available at https://github.com/sparisi/cbet/.
****
SimiGrad: Fine-Grained Adaptive Batching for Large Scale Training using Gradient Similarity Measurement
Heyang Qin, Samyam Rajbhandari, Olatunji Ruwase, Feng Yan, Lei Yang, Yuxiong He
Large scale training requires massive parallelism to finish the training within a reasonable amount of time. To support massive parallelism, large batch training is the key enabler but often at the cost of generalization performance. Existing works explore adaptive batching or hand-tuned static large batching, in order to strike a balance between the computational efficiency and the performance. However, these methods can provide only coarse-grained adaption (e.g., at a epoch level) due to the intrinsic expensive calculation or hand tuning requirements. In this paper, we propose a fully automated and lightweight adaptive batching methodology to enable fine-grained batch size adaption (e.g., at a mini-batch level) that can achieve state-of-the-art performance with record breaking batch sizes. The core component of our method is a lightweight yet efficient representation of the critical gradient noise information. We open-source the proposed methodology by providing a plugin tool that supports mainstream machine learning frameworks. Extensive evaluations on popular benchmarks (e.g., CIFAR10, ImageNet, and

BERT-Large) demonstrate that the proposed methodology outperforms state-of-the-art methodologies using adaptive batching approaches or hand-tuned static strategies in both performance and batch size. Particularly, we achieve a new state-of-the-art batch size of 78k in BERT-Large pretraining with SQuAD score 90.69 compared to 90.58 reported in previous state-of-the-art with 59k batch size.
****
Variational Inference for Continuous-Time Switching Dynamical Systems
Lukas Köhs, Bastian Alt, Heinz Koeppl
Switching dynamical systems provide a powerful, interpretable modeling framework for inference in time-series data in, e.g., the natural sciences or engineering applications. Since many areas, such as biology or discrete-event systems, are naturally described in continuous time, we present a model based on a Markov jump process modulating a subordinated diffusion process. We provide the exact evolution equations for the prior and posterior marginal densities, the direct solutions of which are however computationally intractable. Therefore, we develop a new continuous-time variational inference algorithm, combining a Gaussian process approximation on the diffusion level with posterior inference for Markov jump processes. By minimizing the path-wise Kullback-Leibler divergence we obtain (i) Bayesian latent state estimates for arbitrary points on the real axis and (ii) point estimates of unknown system parameters, utilizing variational expectation maximization. We extensively evaluate our algorithm under the model assumption and for real-world examples.
****
Implicit Regularization in Matrix Sensing via Mirror Descent
Fan Wu, Patrick Rebeschini
We study discrete-time mirror descent applied to the unregularized empirical risk in matrix sensing. In both the general case of rectangular matrices and the particular case of positive semidefinite matrices, a simple potential-based analysis in terms of the Bregman divergence allows us to establish convergence of mirror descent---with different choices of the mirror maps---to a matrix that, among all global minimizers of the empirical risk, minimizes a quantity explicitly related to the nuclear norm, the Frobenius norm, and the von Neumann entropy. In both cases, this characterization implies that mirror descent, a first-order algorithm minimizing the unregularized empirical risk, recovers low-rank matrices under the same set of assumptions that are sufficient to guarantee recovery for nuclear-norm minimization. When the sensing matrices are symmetric and commute, we show that gradient descent with full-rank factorized parametrization is a first-order approximation to mirror descent, in which case we obtain an explicit characterization of the implicit bias of gradient flow as a by-product.
****
STORM+: Fully Adaptive SGD with Recursive Momentum for Nonconvex Optimization
Kfir Levy, 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-authors prior to requesting a name change in the electronic proceedings.
****
Skipping the Frame-Level: Event-Based Piano Transcription With Neural Semi-CRFs
Yujia Yan, Frank Cwitkowitz, Zhiyao Duan
Piano transcription systems are typically optimized to estimate pitch activity at each frame of audio. They are often followed by carefully designed heuristics and post-processing algorithms to estimate note events from the frame-level predictions. Recent methods have also framed piano transcription as a multi-task learning problem, where the activation of different stages of a note event are estimated independently. These practices are not well aligned with the desired outcome of the task, which is the specification of note intervals as holistic events, rather than the aggregation of disjoint observations. In this work, we propose a novel formulation of piano transcription, which is optimized to directly predict note events. Our method is based on Semi-Markov Conditional Random Fields (semi-CRF), which produce scores for intervals rather than individual frames. When

formulating piano transcription in this way, we eliminate the need to rely on di sjoint frame-level estimates for different stages of a note event. We conduct ex periments on the MAESTRO dataset and demonstrate that the proposed model surpass es the current state-of-the-art for piano transcription. Our results suggest tha t the semi-CRF output layer, while still quadratic in complexity, is a simple, f ast and well-performing solution for event-based prediction, and may lead to sim ilar success in other areas which currently rely on frame-level estimates.
****

Deep Learning on a Data Diet: Finding Important Examples Early in Training
Mansheej Paul, Surya Ganguli, Gintare Karolina Dziugaite
Recent success in deep learning has partially been driven by training increasing ly overparametrized networks on ever larger datasets. It is therefore natural to ask: how much of the data is superfluous, which examples are important for gene ralization, and how do we find them? In this work, we make the striking observat ion that, in standard vision datasets, simple scores averaged over several weigh t initializations can be used to identify important examples very early in train ing. We propose two such scores—the Gradient Normed (GraNd) and the Error L2-Nor m (EL2N) scores—and demonstrate their efficacy on a range of architectures and d atasets by pruning significant fractions of training data without sacrificing te st accuracy. In fact, using EL2N scores calculated a few epochs into training, w e can prune half of the CIFAR10 training set while slightly improving test accur acy. Furthermore, for a given dataset, EL2N scores from one architecture or hype rparameter configuration generalize to other configurations. Compared to recent work that prunes data by discarding examples that are rarely forgotten over the course of training, our scores use only local information early in training. We also use our scores to detect noisy examples and study training dynamics through the lens of important examples—we investigate how the data distribution shapes the loss surface and identify subspaces of the model’s data representation that are relatively stable over training.
****

BNS: Building Network Structures Dynamically for Continual Learning
Qi Qin, Wenpeng Hu, Han Peng, Dongyan Zhao, Bing Liu
Continual learning (CL) of a sequence of tasks is often accompanied with the cat astrophic forgetting(CF) problem. Existing research has achieved remarkable resu lts in overcoming CF, especially for task continual learning. However, limited w ork has been done to achieve another important goal of CL,knowledge transfer.In this paper, we propose a technique (called BNS) to do both. The novelty of BNS is that it dynamically builds a network to learn each new task to overcome CF an d to transfer knowledge across tasks at the same time. Experimental results show that when the tasks are different (with little shared knowledge), BNS can alrea dy outperform the state-of-the-art baselines. When the tasks are similar and hav e shared knowledge, BNS outperforms the baselines substantially by a large margi n due to its knowledge transfer capability.
****

Auditing Black-Box Prediction Models for Data Minimization Compliance
Bashir Rastegarpanah, Krishna Gummadi, Mark Crovella
In this paper, we focus on auditing black-box prediction models for compliance w ith the GDPR’s data minimization principle. This principle restricts prediction models to use the minimal information that is necessary for performing the task at hand. Given the challenge of the black-box setting, our key idea is to check if each of the prediction model’s input features is individually necessary by as signing it some constant value (i.e., applying a simple imputation) across all p rediction instances, and measuring the extent to which the model outcomes would change. We introduce a metric for data minimization that is based on model insta bility under simple imputations. We extend the applicability of this metric from a finite sample model to a distributional setting by introducing a probabilisti c data minimization guarantee, which we derive using a Bayesian approach. Furthe rmore, we address the auditing problem under a constraint on the number of queri es to the prediction system. We formulate the problem of allocating a budget of system queries to feasible simple imputations (for investigating model instabili

ty) as a multi-armed bandit framework with probabilistic success metrics. We define two bandit problems for providing a probabilistic data minimization guarantee at a given confidence level: a decision problem given a data minimization level, and a measurement problem given a fixed query budget. We design efficient algorithms for these auditing problems using novel exploration strategies that expand classical bandit strategies. Our experiments with real-world prediction systems show that our auditing algorithms significantly outperform simpler benchmarks in both measurement and decision problems.
****

Dueling Bandits with Team Comparisons
Lee Cohen, Ulrike Schmidt-Kraepelin, Yishay Mansour
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

Meta Internal Learning
Raphael Bensadoun, Shir Gur, 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.
****

Uniform Convergence of Interpolators: Gaussian Width, Norm Bounds and Benign Overfitting
Frederic Koehler, Lijia Zhou, Danica J. Sutherland, Nathan Srebro
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

Adaptive wavelet distillation from neural networks through interpretations
Wooseok Ha, Chandan Singh, Francois Lanusse, Srigokul Upadhyayula, Bin Yu
Recent deep-learning models have achieved impressive prediction performance, but often sacrifice interpretability and computational efficiency. Interpretability is crucial in many disciplines, such as science and medicine, where models must be carefully vetted or where interpretation is the goal itself. Moreover, interpretable models are concise and often yield computational efficiency. Here, we propose adaptive wavelet distillation (AWD), a method which aims to distill information from a trained neural network into a wavelet transform. Specifically, AWD penalizes feature attributions of a neural network in the wavelet domain to learn an effective multi-resolution wavelet transform. The resulting model is highly predictive, concise, computationally efficient, and has properties (such as a multi-scale structure) which make it easy to interpret. In close collaboration with domain experts, we showcase how AWD addresses challenges in two real-world settings: cosmological parameter inference and molecular-partner prediction. In both cases, AWD yields a scientifically interpretable and concise model which gives predictive performance better than state-of-the-art neural networks. Moreover, AWD identifies predictive features that are scientifically meaningful in the context of respective domains. All code and models are released in a full-fledged package available on Github.
****

Generative Occupancy Fields for 3D Surface-Aware Image Synthesis
Xudong XU, Xingang Pan, Dahua Lin, Bo Dai
The advent of generative radiance fields has significantly promoted the development of 3D-aware image synthesis. The cumulative rendering process in radiance fields makes training these generative models much easier since gradients are distributed over the entire volume, but leads to diffused object surfaces. In the meantime, compared to radiance fields occupancy representations could inherently ensure deterministic surfaces. However, if we directly apply occupancy representa

tions to generative models, during training they will only receive sparse gradie nts located on object surfaces and eventually suffer from the convergence proble m. In this paper, we propose Generative Occupancy Fields (GOF), a novel model ba sed on generative radiance fields that can learn compact object surfaces without impeding its training convergence. The key insight of GOF is a dedicated transi tion from the cumulative rendering in radiance fields to rendering with only the surface points as the learned surface gets more and more accurate. In this way, GOF combines the merits of two representations in a unified framework. In pract ice, the training-time transition of start from radiance fields and march to occ upancy representations is achieved in GOF by gradually shrinking the sampling re gion in its rendering process from the entire volume to a minimal neighboring re gion around the surface. Through comprehensive experiments on multiple datasets, we demonstrate that GOF can synthesize high-quality images with 3D consistency and simultaneously learn compact and smooth object surfaces. Our code is availab le at https://github.com/SheldonTsui/GOF_NeurIPS2021.
****

Relaxed Marginal Consistency for Differentially Private Query Answering

Ryan McKenna, Siddhant Pradhan, Daniel R. Sheldon, Gerome Miklau

Many differentially private algorithms for answering database queries involve as tep that reconstructs a discrete data distribution from noisy measurements. This provides consistent query answers and reduces error, but often requires space th atgrows exponentially with dimension. PRIVATE-PGM is a recent approach that uses graphical models to represent the data distribution, with complexity proportiona l tothat of exact marginal inference in a graphical model with structure determi ned bythe co-occurrence of variables in the noisy measurements. PRIVATE-PGM is h ighlyscalable for sparse measurements, but may fail to run in high dimensions wi th densemeasurements. We overcome the main scalability limitation of PRIVATE-PGM through a principled approach that relaxes consistency constraints in the estima tionobjective. Our new approach works with many existing private query answering algorithms and improves scalability or accuracy with no privacy cost.
****

Local policy search with Bayesian optimization

Sarah Müller, Alexander von Rohr, Sebastian Trimpe

Reinforcement learning (RL) aims to find an optimal policy by interaction with a n environment. Consequently, learning complex behavior requires a vast number of samples, which can be prohibitive in practice. Nevertheless, instead of systema tically reasoning and actively choosing informative samples, policy gradients fo r local search are often obtained from random perturbations. These random sample s yield high variance estimates and hence are sub-optimal in terms of sample com plexity. Actively selecting informative samples is at the core of Bayesian optim ization, which constructs a probabilistic surrogate of the objective from past s amples to reason about informative subsequent ones. In this paper, we propose to join both worlds. We develop an algorithm utilizing a probabilistic model of th e objective function and its gradient. Based on the model, the algorithm decides where to query a noisy zeroth-order oracle to improve the gradient estimates. T he resulting algorithm is a novel type of policy search method, which we compare to existing black-box algorithms. The comparison reveals improved sample comple xity and reduced variance in extensive empirical evaluations on synthetic object ives. Further, we highlight the benefits of active sampling on popular RL benchm arks.
****

DominoSearch: Find layer-wise fine-grained N:M sparse schemes from dense neural networks

Wei Sun, Aojun Zhou, Sander Stuijk, Rob Wijnhoven, Andrew Oakleigh Nelson, hongs heng Li, Henk Corporaal

Neural pruning is a widely-used compression technique for Deep Neural Networks (DNNs). Recent innovations in Hardware Architectures (e.g. Nvidia Ampere Sparse T ensor Core) and N:M fine-grained Sparse Neural Network algorithms (i.e. every M-weights contains N non-zero values) reveal a promising research line of neural p runing. However, the existing N:M algorithms only address the challenge of how t

o train N:M sparse neural networks in a uniform fashion (i.e. every layer has the same N:M sparsity) and suffer from a significant accuracy drop for high sparsity (i.e. when sparsity > 80\%). To tackle this problem, we present a novel technique -- \textbf{\textit{DominoSearch}} to find mixed N:M sparsity schemes from pre-trained dense deep neural networks to achieve higher accuracy than the uniform-sparsity scheme with equivalent complexity constraints (e.g. model size or FLOPs). For instance, for the same model size with 2.1M parameters (87.5\% sparsity), our layer-wise N:M sparse ResNet18 outperforms its uniform counterpart by 2.1\% top-1 accuracy, on the large-scale ImageNet dataset. For the same computational complexity of 227M FLOPs, our layer-wise sparse ResNet18 outperforms the uniform one by 1.3\% top-1 accuracy. Furthermore, our layer-wise fine-grained N:M sparse ResNet50 achieves 76.7\% top-1 accuracy with 5.0M parameters. {This is competitive to the results achieved by layer-wise unstructured sparsity} that is believed to be the upper-bound of Neural Network pruning with respect to the accuracy-sparsity trade-off. We believe that our work can build a strong baseline for further sparse DNN research and encourage future hardware-algorithm co-design work. Our code and models are publicly available at \url{https://github.com/NM-sparsity/DominoSearch}.
****

Techniques for Symbol Grounding with SATNet

Sever Topan, David Rolnick, Xujie Si

Many experts argue that the future of artificial intelligence is limited by the field’s ability to integrate symbolic logical reasoning into deep learning architectures. The recently proposed differentiable MAXSAT solver, SATNet, was a breakthrough in its capacity to integrate with a traditional neural network and solve visual reasoning problems. For instance, it can learn the rules of Sudoku purely from image examples. Despite its success, SATNet was shown to succumb to a key challenge in neurosymbolic systems known as the Symbol Grounding Problem: the inability to map visual inputs to symbolic variables without explicit supervision ("label leakage"). In this work, we present a self-supervised pre-training pipeline that enables SATNet to overcome this limitation, thus broadening the class of problems that SATNet architectures can solve to include datasets where no intermediary labels are available at all. We demonstrate that our method allows SATNet to attain full accuracy even with a harder problem setup that prevents any label leakage. We additionally introduce a proofreading method that further improves the performance of SATNet architectures, beating the state-of-the-art on Visual Sudoku.
****

Object DGCNN: 3D Object Detection using Dynamic Graphs

Yue Wang, Justin M. Solomon

3D object detection often involves complicated training and testing pipelines, which require substantial domain knowledge about individual datasets. Inspired by recent non-maximum suppression-free 2D object detection models, we propose a 3D object detection architecture on point clouds. Our method models 3D object detection as message passing on a dynamic graph, generalizing the DGCNN framework to predict a set of objects. In our construction, we remove the necessity of post-processing via object confidence aggregation or non-maximum suppression. To facilitate object detection from sparse point clouds, we also propose a set-to-set distillation approach customized to 3D detection. This approach aligns the outputs of the teacher model and the student model in a permutation-invariant fashion, significantly simplifying knowledge distillation for the 3D detection task. Our method achieves state-of-the-art performance on autonomous driving benchmarks. We also provide abundant analysis of the detection model and distillation framework.
****

Safe Policy Optimization with Local Generalized Linear Function Approximations

Akifumi Wachi, Yunyue Wei, Yanan Sui

Safe exploration is a key to applying reinforcement learning (RL) in safety-critical systems. Existing safe exploration methods guaranteed safety under the assumption of regularity, and it has been difficult to apply them to large-scale rea

l problems. We propose a novel algorithm, SPO-LF, that optimizes an agent's poli
cy while learning the relation between a locally available feature obtained by s
ensors and environmental reward/safety using generalized linear function approxi
mations. We provide theoretical guarantees on its safety and optimality. We expe
rimentally show that our algorithm is 1) more efficient in terms of sample compl
exity and computational cost and 2) more applicable to large-scale problems than
 previous safe RL methods with theoretical guarantees, and 3) comparably sample-
efficient and safer compared with existing advanced deep RL methods with safety
constraints.
****
Symplectic Adjoint Method for Exact Gradient of Neural ODE with Minimal Memory
Takashi Matsubara, Yuto Miyatake, Takaharu Yaguchi
A neural network model of a differential equation, namely neural ODE, has enable
d the learning of continuous-time dynamical systems and probabilistic distributi
ons with high accuracy. The neural ODE uses the same network repeatedly during a
 numerical integration. The memory consumption of the backpropagation algorithm
is proportional to the number of uses times the network size. This is true even
if a checkpointing scheme divides the computation graph into sub-graphs. Otherwi
se, the adjoint method obtains a gradient by a numerical integration backward in
 time. Although this method consumes memory only for a single network use, it re
quires high computational cost to suppress numerical errors. This study proposes
 the symplectic adjoint method, which is an adjoint method solved by a symplecti
c integrator. The symplectic adjoint method obtains the exact gradient (up to ro
unding error) with memory proportional to the number of uses plus the network si
ze. The experimental results demonstrate that the symplectic adjoint method cons
umes much less memory than the naive backpropagation algorithm and checkpointing
 schemes, performs faster than the adjoint method, and is more robust to roundin
g errors.
****
Exponential Separation between Two Learning Models and Adversarial Robustness
Grzegorz Gluch, Ruediger Urbanke
Requests for name changes in the electronic proceedings will be accepted with 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 balancing principle for parameter choice in distance-regularized domain adap
tation
Werner Zellinger, Natalia Shepeleva, Marius-Constantin Dinu, Hamid Eghbal-zadeh,
 Hoan Duc Nguyen, Bernhard Nessler, Sergei Pereverzyev, Bernhard A. Moser
We address the unsolved algorithm design problem of choosing a justified regular
ization parameter in unsupervised domain adaptation. This problem is intriguing
as no labels are available in the target domain. Our approach starts with the ob
servation that the widely-used method of minimizing the source error, penalized
by a distance measure between source and target feature representations, shares
characteristics with regularized ill-posed inverse problems. Regularization para
meters in inverse problems are optimally chosen by the fundamental principle of
balancing approximation and sampling errors. We use this principle to balance le
arning errors and domain distance in a target error bound. As a result, we obtai
n a theoretically justified rule for the choice of the regularization parameter.
 In contrast to the state of the art, our approach allows source and target dist
ributions with disjoint supports. An empirical comparative study on benchmark da
tasets underpins the performance of our approach.
****
Gaussian Kernel Mixture Network for Single Image Defocus Deblurring
Yuhui Quan, Zicong Wu, Hui Ji
Defocus blur is one kind of blur effects often seen in images, which is challeng
ing to remove due to its spatially variant amount. This paper presents an end-t
o-end deep learning approach for removing defocus blur from a single image, so a
s to have an all-in-focus image for consequent vision tasks. First, a pixel-wis

e Gaussian kernel mixture (GKM) model is proposed for representing spatially va
riant defocus blur kernels in an efficient linear parametric form, with higher a
ccuracy than existing models. Then, a deep neural network called GKMNet is devel
oped by unrolling a fixed-point iteration of the GKM-based deblurring. The GKMN
et is built on a lightweight scale-recurrent architecture, with a scale-recurren
t attention module for estimating the mixing coefficients in GKM for defocus deb
lurring. Extensive experiments show that the GKMNet not only noticeably outperfo
rms existing defocus deblurring methods, but also has its advantages in terms of
 model complexity and computational efficiency.
****
Cockpit: A Practical Debugging Tool for the Training of Deep Neural Networks
Frank Schneider, Felix Dangel, Philipp Hennig
When engineers train deep learning models, they are very much "flying blind". Co
mmonly used methods for real-time training diagnostics, such as monitoring the t
rain/test loss, are limited. Assessing a network's training process solely throu
gh these performance indicators is akin to debugging software without access to
internal states through a debugger. To address this, we present Cockpit, a colle
ction of instruments that enable a closer look into the inner workings of a lear
ning machine, and a more informative and meaningful status report for practition
ers. It facilitates the identification of learning phases and failure modes, lik
e ill-chosen hyperparameters. These instruments leverage novel higher-order info
rmation about the gradient distribution and curvature, which has only recently b
ecome efficiently accessible. We believe that such a debugging tool, which we op
en-source for PyTorch, is a valuable help in troubleshooting the training proces
s. By revealing new insights, it also more generally contributes to explainabili
ty and interpretability of deep nets.
****
MEST: Accurate and Fast Memory-Economic Sparse Training Framework on the Edge
Geng Yuan, Xiaolong Ma, Wei Niu, Zhengang Li, Zhenglun Kong, Ning Liu, Yifan Gon
g, Zheng Zhan, Chaoyang He, Qing Jin, Siyue Wang, Minghai Qin, Bin Ren, Yanzhi W
ang, Sijia Liu, Xue Lin
Recently, a new trend of exploring sparsity for accelerating neural network trai
ning has emerged, embracing the paradigm of training on the edge. This paper pro
poses a novel Memory-Economic Sparse Training (MEST) framework targeting for acc
urate and fast execution on edge devices. The proposed MEST framework consists o
f enhancements by Elastic Mutation (EM) and Soft Memory Bound (&S) that ensure s
uperior accuracy at high sparsity ratios. Different from the existing works for
sparse training, this current work reveals the importance of sparsity schemes on
 the performance of sparse training in terms of accuracy as well as training spe
ed on real edge devices. On top of that, the paper proposes to employ data effic
iency for further acceleration of sparse training. Our results suggest that unfo
rgettable examples can be identified in-situ even during the dynamic exploration
 of sparsity masks in the sparse training process, and therefore can be removed
for further training speedup on edge devices. Comparing with state-of-the-art (S
OTA) works on accuracy, our MEST increases Top-1 accuracy significantly on Image
Net when using the same unstructured sparsity scheme. Systematical evaluation on
 accuracy, training speed, and memory footprint are conducted, where the propose
d MEST framework consistently outperforms representative SOTA works. A reviewer
strongly against our work based on his false assumptions and misunderstandings.
On top of the previous submission, we employ data efficiency for further acceler
ation of sparse training. And we explore the impact of model sparsity, sparsity
schemes, and sparse training algorithms on the number of removable training exam
ples. Our codes are publicly available at: https://github.com/boone891214/MEST.
****
Precise characterization of the prior predictive distribution of deep ReLU netwo
rks
Lorenzo Noci, Gregor Bachmann, Kevin Roth, Sebastian Nowozin, Thomas Hofmann
Recent works on Bayesian neural networks (BNNs) have highlighted the need to bet
ter understand the implications of using Gaussian priors in combination with the
 compositional structure of the network architecture. Similar in spirit to the k

ind of analysis that has been developed to devise better initialization schemes for neural networks (cf. He- or Xavier initialization), we derive a precise characterization of the prior predictive distribution of finite-width ReLU networks with Gaussian weights.While theoretical results have been obtained for their heavy-tailedness,the full characterization of the prior predictive distribution (i.e. its density, CDF and moments), remained unknown prior to this work. Our analysis, based on the Meijer-G function, allows us to quantify the influence of architectural choices such as the width or depth of the network on the resulting shape of the prior predictive distribution. We also formally connect our results to previous work in the infinite width setting, demonstrating that the moments of the distribution converge to those of a normal log-normal mixture in the infinite depth limit. Finally, our results provide valuable guidance on prior design: for instance, controlling the predictive variance with depth- and width-informed priors on the weights of the network.
****
RED : Looking for Redundancies for Data-FreeStructured Compression of Deep Neural Networks
Edouard YVINEC, Arnaud Dapogny, Matthieu Cord, Kevin Bailly
Deep Neural Networks (DNNs) are ubiquitous in today's computer vision landscape, despite involving considerable computational costs. The mainstream approaches for runtime acceleration consist in pruning connections (unstructured pruning) or, better, filters (structured pruning), both often requiring data to retrain the model. In this paper, we present RED, a data-free, unified approach to tackle structured pruning. First, we propose a novel adaptive hashing of the scalar DNN weight distribution densities to increase the number of identical neurons represented by their weight vectors. Second, we prune the network by merging redundant neurons based on their relative similarities, as defined by their distance. Third, we propose a novel uneven depthwise separation technique to further prune convolutional layers. We demonstrate through a large variety of benchmarks that RED largely outperforms other data-free pruning methods, often reaching performance similar to unconstrained, data-driven methods.
****
TestRank: Bringing Order into Unlabeled Test Instances for Deep Learning Tasks
YU LI, Min LI, Qiuxia LAI, Yannan Liu, Qiang Xu
Deep learning (DL) systems are notoriously difficult to test and debug due to the lack of correctness proof and the huge test input space to cover. Given the ubiquitous unlabeled test data and high labeling cost, in this paper, we propose a novel test prioritization technique, namely TestRank, which aims at revealing more model failures with less labeling effort. TestRank brings order into the unlabeled test data according to their likelihood of being a failure, i.e., their failure-revealing capabilities. Different from existing solutions, TestRank leverages both intrinsic and contextual attributes of the unlabeled test data when prioritizing them. To be specific, we first build a similarity graph on both unlabeled test samples and labeled samples (e.g., training or previously labeled test samples). Then, we conduct graph-based semi-supervised learning to extract contextual features from the correctness of similar labeled samples. For a particular test instance, the contextual features extracted with the graph neural network and the intrinsic features obtained with the DL model itself are combined to predict its failure-revealing capability. Finally, TestRank prioritizes unlabeled test inputs in descending order of the above probability value. We evaluate TestRank on three popular image classification datasets, and results show that TestRank significantly outperforms existing test prioritization techniques.
****
Large Scale Learning on Non-Homophilous Graphs: New Benchmarks and Strong Simple Methods
Derek Lim, Felix Hohne, Xiuyu Li, Sijia Linda Huang, Vaishnavi Gupta, Omkar Bhalerao, Ser Nam Lim
Many widely used datasets for graph machine learning tasks have generally been homophilous, where nodes with similar labels connect to each other. Recently, new Graph Neural Networks (GNNs) have been developed that move beyond the homophily

regime; however, their evaluation has often been conducted on small graphs with limited application domains. We collect and introduce diverse non-homophilous datasets from a variety of application areas that have up to 384x more nodes and 1398x more edges than prior datasets. We further show that existing scalable graph learning and graph minibatching techniques lead to performance degradation on these non-homophilous datasets, thus highlighting the need for further work on scalable non-homophilous methods. To address these concerns, we introduce LINKX --- a strong simple method that admits straightforward minibatch training and inference. Extensive experimental results with representative simple methods and GNNs across our proposed datasets show that LINKX achieves state-of-the-art performance for learning on non-homophilous graphs. Our codes and data are available at https://github.com/CUAI/Non-Homophily-Large-Scale.
****
Reinforcement Learning based Disease Progression Model for Alzheimer’s Disease
Krishnakant Saboo, Anirudh Choudhary, Yurui Cao, Gregory Worrell, David Jones, Ravishankar Iyer
We model Alzheimer’s disease (AD) progression by combining differential equations (DEs) and reinforcement learning (RL) with domain knowledge. DEs provide relationships between some, but not all, factors relevant to AD. We assume that the missing relationships must satisfy general criteria about the working of the brain, for e.g., maximizing cognition while minimizing the cost of supporting cognition. This allows us to extract the missing relationships by using RL to optimize an objective (reward) function that captures the above criteria. We use our model consisting of DEs (as a simulator) and the trained RL agent to predict individualized 10-year AD progression using baseline (year 0) features on synthetic and real data. The model was comparable or better at predicting 10-year cognition trajectories than state-of-the-art learning-based models. Our interpretable model demonstrated, and provided insights into, "recovery/compensatory" processes that mitigate the effect of AD, even though those processes were not explicitly encoded in the model. Our framework combines DEs with RL for modelling AD progression and has broad applicability for understanding other neurological disorders.
****
Catch-A-Waveform: Learning to Generate Audio from a Single Short Example
Gal Greshler, Tamar Shaham, Tomer Michaeli
Models for audio generation are typically trained on hours of recordings. Here, we illustrate that capturing the essence of an audio source is typically possible from as little as a few tens of seconds from a single training signal. Specifically, we present a GAN-based generative model that can be trained on one short audio signal from any domain (e.g. speech, music, etc.) and does not require pre-training or any other form of external supervision. Once trained, our model can generate random samples of arbitrary duration that maintain semantic similarity to the training waveform, yet exhibit new compositions of its audio primitives. This enables a long line of interesting applications, including generating new jazz improvisations or new a-cappella rap variants based on a single short example, producing coherent modifications to famous songs (e.g. adding a new verse to a Beatles song based solely on the original recording), filling-in of missing parts (inpainting), extending the bandwidth of a speech signal (super-resolution), and enhancing old recordings without access to any clean training example. We show that in all cases, no more than 20 seconds of training audio commonly suffice for our model to achieve state-of-the-art results. This is despite its complete lack of prior knowledge about the nature of audio signals in general.
****
Explanation-based Data Augmentation for Image Classification
Sandareka Wickramanayake, Wynne Hsu, Mong Li Lee
Existing works have generated explanations for deep neural network decisions to provide insights into model behavior. We observe that these explanations can also be used to identify concepts that caused misclassifications. This allows us to understand the possible limitations of the dataset used to train the model, particularly the under-represented regions in the dataset. This work proposes a framework that utilizes concept-based explanations to automatically augment the dat

aset with new images that can cover these under-represented regions to improve the model performance. The framework is able to use the explanations generated by both interpretable classifiers and post-hoc explanations from black-box classifiers. Experiment results demonstrate that the proposed approach improves the accuracy of classifiers compared to state-of-the-art augmentation strategies.
****
Data-Efficient GAN Training Beyond (Just) Augmentations: A Lottery Ticket Perspective
Tianlong Chen, Yu Cheng, Zhe Gan, Jingjing Liu, Zhangyang Wang
Training generative adversarial networks (GANs) with limited real image data generally results in deteriorated performance and collapsed models. To conquer this challenge, we are inspired by the latest observation, that one can discover independently trainable and highly sparse subnetworks (a.k.a., lottery tickets) from GANs. Treating this as an inductive prior, we suggest a brand-new angle towards data-efficient GAN training: by first identifying the lottery ticket from the original GAN using the small training set of real images; and then focusing on training that sparse subnetwork by re-using the same set. We find our coordinated framework to offer orthogonal gains to existing real image data augmentation methods, and we additionally present a new feature-level augmentation that can be applied together with them. Comprehensive experiments endorse the effectiveness of our proposed framework, across various GAN architectures (SNGAN, BigGAN, and StyleGAN-V2) and diverse datasets (CIFAR-10, CIFAR-100, Tiny-ImageNet, ImageNet, and multiple few-shot generation datasets). Codes are available at: https://github.com/VITA-Group/Ultra-Data-Efficient-GAN-Training.
****
When Are Solutions Connected in Deep Networks?
Quynh N. Nguyen, Pierre Bréchet, Marco Mondelli
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
TOHAN: A One-step Approach towards Few-shot Hypothesis Adaptation
Haoang Chi, Feng Liu, Wenjing Yang, Long Lan, Tongliang Liu, Bo Han, William Cheung, James Kwok
In few-shot domain adaptation (FDA), classifiers for the target domain are trained with \emph{accessible} labeled data in the source domain (SD) and few labeled data in the target domain (TD). However, data usually contain private information in the current era, e.g., data distributed on personal phones. Thus, the private data will be leaked if we directly access data in SD to train a target-domain classifier (required by FDA methods). In this paper, to prevent privacy leakage in SD, we consider a very challenging problem setting, where the classifier for the TD has to be trained using few labeled target data and a well-trained SD classifier, named few-shot hypothesis adaptation (FHA). In FHA, we cannot access data in SD, as a result, the private information in SD will be protected well. To this end, we propose a target-oriented hypothesis adaptation network (TOHAN) to solve the FHA problem, where we generate highly-compatible unlabeled data (i.e., an intermediate domain) to help train a target-domain classifier. TOHAN maintains two deep networks simultaneously, in which one focuses on learning an intermediate domain and the other takes care of the intermediate-to-target distributional adaptation and the target-risk minimization. Experimental results show that TOHAN outperforms competitive baselines significantly.
****
Learning Graph Cellular Automata
Daniele Grattarola, Lorenzo Livi, Cesare Alippi
Cellular automata (CA) are a class of computational models that exhibit rich dynamics emerging from the local interaction of cells arranged in a regular lattice. In this work we focus on a generalised version of typical CA, called graph cellular automata (GCA), in which the lattice structure is replaced by an arbitrary graph. In particular, we extend previous work that used convolutional neural ne

tworks to learn the transition rule of conventional CA and we use graph neural networks to learn a variety of transition rules for GCA. First, we present a general-purpose architecture for learning GCA, and we show that it can represent any arbitrary GCA with finite and discrete state space. Then, we test our approach on three different tasks: 1) learning the transition rule of a GCA on a Voronoi tessellation; 2) imitating the behaviour of a group of flocking agents; 3) learning a rule that converges to a desired target state.
****

Efficient Online Estimation of Causal Effects by Deciding What to Observe
Shantanu Gupta, Zachary Lipton, David Childers
Researchers often face data fusion problems, where multiple data sources are available, each capturing a distinct subset of variables. While problem formulations typically take the data as given, in practice, data acquisition can be an ongoing process. In this paper, we introduce the problem of deciding, at each time, which data source to sample from. Our goal is to estimate a given functional of the parameters of a probabilistic model as efficiently as possible. We propose online moment selection (OMS), a framework in which structural assumptions are encoded as moment conditions. The optimal action at each step depends, in part, on the very moments that identify the functional of interest. Our algorithms balance exploration with choosing the best action as suggested by estimated moments. We propose two selection strategies: (1) explore-then-commit (ETC) and (2) explore-then-greedy (ETG), proving that both achieve zero asymptotic regret as assessed by MSE. We instantiate our setup for average treatment effect estimation, where structural assumptions are given by a causal graph and data sources include subsets of mediators, confounders, and instrumental variables.
****

Perturbation Theory for the Information Bottleneck
Vudtiwat Ngampruetikorn, David J. Schwab
Extracting relevant information from data is crucial for all forms of learning. The information bottleneck (IB) method formalizes this, offering a mathematically precise and conceptually appealing framework for understanding learning phenomena. However the nonlinearity of the IB problem makes it computationally expensive and analytically intractable in general. Here we derive a perturbation theory for the IB method and report the first complete characterization of the learning onset, the limit of maximum relevant information per bit extracted from data. We test our results on synthetic probability distributions, finding good agreement with the exact numerical solution near the onset of learning. We explore the difference and subtleties in our derivation and previous attempts at deriving a perturbation theory for the learning onset and attribute the discrepancy to a flawed assumption. Our work also provides a fresh perspective on the intimate relationship between the IB method and the strong data processing inequality.
****

Deconvolutional Networks on Graph Data
Jia Li, Jiajin Li, Yang Liu, Jianwei Yu, Yueting Li, Hong Cheng
In this paper, we consider an inverse problem in graph learning domain -- "given the graph representations smoothed by Graph Convolutional Network (GCN), how can we reconstruct the input graph signal?" We propose Graph Deconvolutional Network (GDN) and motivate the design of GDN via a combination of inverse filters in spectral domain and de-noising layers in wavelet domain, as the inverse operation results in a high frequency amplifier and may amplify the noise. We demonstrate the effectiveness of the proposed method on several tasks including graph feature imputation and graph structure generation.
****

Variational Multi-Task Learning with Gumbel-Softmax Priors
Jiayi Shen, Xiantong Zhen, Marcel Worring, Ling Shao
Multi-task learning aims to explore task relatedness to improve individual tasks, which is of particular significance in the challenging scenario that only limited data is available for each task. To tackle this challenge, we propose variational multi-task learning (VMTL), a general probabilistic inference framework for learning multiple related tasks. We cast multi-task learning as a variational

Bayesian inference problem, in which task relatedness is explored in a unified m
anner by specifying priors. To incorporate shared knowledge into each task, we d
esign the prior of a task to be a learnable mixture of the variational posterior
s of other related tasks, which is learned by the Gumbel-Softmax technique. In c
ontrast to previous methods, our VMTL can exploit task relatedness for both repr
esentations and classifiers in a principled way by jointly inferring their poste
riors. This enables individual tasks to fully leverage inductive biases provided
 by related tasks, therefore improving the overall performance of all tasks. Exp
erimental results demonstrate that the proposed VMTL is able to effectively tack
le a variety of challenging multi-task learning settings with limited training d
ata for both classification and regression. Our method consistently surpasses pr
evious methods, including strong Bayesian approaches, and achieves state-of-the-
art performance on five benchmark datasets.
****
Accelerating Quadratic Optimization with Reinforcement Learning
Jeffrey Ichnowski, Paras Jain, Bartolomeo Stellato, Goran Banjac, Michael Luo, F
rancesco Borrelli, Joseph E. Gonzalez, Ion Stoica, Ken Goldberg
First-order methods for quadratic optimization such as OSQP are widely used for
large-scale machine learning and embedded optimal control, where many related pr
oblems must be rapidly solved. These methods face two persistent challenges: man
ual hyperparameter tuning and convergence time to high-accuracy solutions. To ad
dress these, we explore how Reinforcement Learning (RL) can learn a policy to tu
ne parameters to accelerate convergence. In experiments with well-known QP bench
marks we find that our RL policy, RLQP, significantly outperforms state-of-the-a
rt QP solvers by up to 3x. RLQP generalizes surprisingly well to previously unse
en problems with varying dimension and structure from different applications, in
cluding the QPLIB, Netlib LP and Maros-M{\'e}sz{\'a}ros problems. Code, models,
and videos are available at https://berkeleyautomation.github.io/rlqp/.
****
Deep Residual Learning in Spiking Neural Networks
Wei Fang, Zhaofei Yu, Yanqi Chen, Tiejun Huang, Timothée Masquelier, Yonghong Ti
an
Deep Spiking Neural Networks (SNNs) present optimization difficulties for gradie
nt-based approaches due to discrete binary activation and complex spatial-tempor
al dynamics. Considering the huge success of ResNet in deep learning, it would
be natural to train deep SNNs with residual learning. Previous Spiking ResNet mi
mics the standard residual block in ANNs and simply replaces ReLU activation lay
ers with spiking neurons, which suffers the degradation problem and can hardly i
mplement residual learning. In this paper, we propose the spike-element-wise (SE
W) ResNet to realize residual learning in deep SNNs. We prove that the SEW ResNe
t can easily implement identity mapping and overcome the vanishing/exploding gra
dient problems of Spiking ResNet. We evaluate our SEW ResNet on ImageNet, DVS Ge
sture, and CIFAR10-DVS datasets, and show that SEW ResNet outperforms the state-
of-the-art directly trained SNNs in both accuracy and time-steps. Moreover, SEW
 ResNet can achieve higher performance by simply adding more layers, providing a
 simple method to train deep SNNs. To our best knowledge, this is the first time
 that directly training deep SNNs with more than 100 layers becomes possible. Ou
r codes are available at https://github.com/fangwei123456/Spike-Element-Wise-Res
Net.
****
Duplex Sequence-to-Sequence Learning for Reversible Machine Translation
Zaixiang Zheng, Hao Zhou, Shujian Huang, Jiajun Chen, Jingjing Xu, Lei Li
Sequence-to-sequence learning naturally has two directions. How to effectively u
tilize supervision signals from both directions? Existing approaches either requ
ire two separate models, or a multitask-learned model but with inferior performa
nce. In this paper, we propose REDER (Reversible Duplex Transformer), a paramete
r-efficient model and apply it to machine translation. Either end of REDER can s
imultaneously input and output a distinct language. Thus REDER enables {\em reve
rsible machine translation} by simply flipping the input and output ends. Experi
ments verify that REDER achieves the first success of reversible machine transla

tion, which helps outperform its multitask-trained baselines by up to 1.3 BLEU.
****

Improved Coresets and Sublinear Algorithms for Power Means in Euclidean Spaces
Vincent Cohen-Addad, David Saulpic, Chris Schwiegelshohn
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

Accelerated Sparse Neural Training: A Provable and Efficient Method to Find N:M Transposable Masks
Itay Hubara, Brian Chmiel, Moshe Island, Ron Banner, Joseph Naor, Daniel Soudry
Unstructured pruning reduces the memory footprint in deep neural networks (DNNs). Recently, researchers proposed different types of structural pruning intending to reduce also the computation complexity. In this work, we first suggest a new measure called mask-diversity which correlates with the expected accuracy of the different types of structural pruning. We focus on the recently suggested N:M fine-grained block sparsity mask, in which for each block of M weights, we have at least N zeros. While N:M fine-grained block sparsity allows acceleration in actual modern hardware, it can be used only to accelerate the inference phase. In order to allow for similar accelerations in the training phase, we suggest a novel transposable fine-grained sparsity mask, where the same mask can be used for both forward and backward passes. Our transposable mask guarantees that both the weight matrix and its transpose follow the same sparsity pattern; thus, the matrix multiplication required for passing the error backward can also be accelerated. We formulate the problem of finding the optimal transposable-mask as a minimum-cost flow problem. Additionally, to speed up the minimum-cost flow computation, we also introduce a fast linear-time approximation that can be used when the masks dynamically change during training. Our experiments suggest a 2x speed-up in the matrix multiplications with no accuracy degradation over vision and language models. Finally, to solve the problem of switching between different structure constraints, we suggest a method to convert a pre-trained model with unstructured sparsity to an N:M fine-grained block sparsity model with little to no training. A reference implementation can be found at https://github.com/papers-submission/structuredtransposablemasks.
****

Learning and Generalization in RNNs
Abhishek Panigrahi, Navin Goyal
Simple recurrent neural networks (RNNs) and their more advanced cousins LSTMs etc. have been very successful in sequence modeling. Their theoretical understanding, however, is lacking and has not kept pace with the progress for feedforward networks, where a reasonably complete understanding in the special case of highly overparametrized one-hidden-layer networks has emerged. In this paper, we make progress towards remedying this situation by proving that RNNs can learn functions of sequences. In contrast to the previous work that could only deal with functions of sequences that are sums of functions of individual tokens in the sequence, we allow general functions. Conceptually and technically, we introduce new ideas which enable us to extract information from the hidden state of the RNN in our proofs---addressing a crucial weakness in previous work. We illustrate our results on some regular language recognition problems.
****

Improving Visual Quality of Image Synthesis by A Token-based Generator with Transformers
Yanhong Zeng, Huan Yang, Hongyang Chao, Jianbo Wang, Jianlong Fu
We present a new perspective of achieving image synthesis by viewing this task as a visual token generation problem. Different from existing paradigms that directly synthesize a full image from a single input (e.g., a latent code), the new formulation enables a flexible local manipulation for different image regions, which makes it possible to learn content-aware and fine-grained style control for image synthesis. Specifically, it takes as input a sequence of latent tokens to

predict the visual tokens for synthesizing an image. Under this perspective, we propose a token-based generator (i.e., TokenGAN). Particularly, the TokenGAN inputs two semantically different visual tokens, i.e., the learned constant content tokens and the style tokens from the latent space. Given a sequence of style tokens, the TokenGAN is able to control the image synthesis by assigning the styles to the content tokens by attention mechanism with a Transformer. We conduct extensive experiments and show that the proposed TokenGAN has achieved state-of-the-art results on several widely-used image synthesis benchmarks, including FFHQ and LSUN CHURCH with different resolutions. In particular, the generator is able to synthesize high-fidelity images with (1024x1024) size, dispensing with convolutions entirely.
****
The Effect of the Intrinsic Dimension on the Generalization of Quadratic Classifiers
Fabian Latorre, Leello Tadesse Dadi, Paul Rolland, Volkan Cevher
It has been recently observed that neural networks, unlike kernel methods, enjoy a reduced sample complexity when the distribution is isotropic (i.e., when the covariance matrix is the identity). We find that this sensitivity to the data distribution is not exclusive to neural networks, and the same phenomenon can be observed on the class of quadratic classifiers (i.e., the sign of a quadratic polynomial) with a nuclear-norm constraint. We demonstrate this by deriving an upper bound on the Rademacher Complexity that depends on two key quantities: (i) the intrinsic dimension, which is a measure of isotropy, and (ii) the largest eigenvalue of the second moment (covariance) matrix of the distribution. Our result improves the dependence on the dimension over the best previously known bound and precisely quantifies the relation between the sample complexity and the level of isotropy of the distribution.
****
DeepReduce: A Sparse-tensor Communication Framework for Federated Deep Learning
Hang Xu, Kelly Kostopoulou, Aritra Dutta, Xin Li, Alexandros Ntoulas, Panos Kalnis
Sparse tensors appear frequently in federated deep learning, either as a direct artifact of the deep neural network’s gradients, or as a result of an explicit sparsification process. Existing communication primitives are agnostic to the peculiarities of deep learning; consequently, they impose unnecessary communication overhead. This paper introduces DeepReduce, a versatile framework for the compressed communication of sparse tensors, tailored to federated deep learning. DeepReduce decomposes sparse tensors into two sets, values and indices, and allows both independent and combined compression of these sets. We support a variety of common compressors, such as Deflate for values, or run-length encoding for indices. We also propose two novel compression schemes that achieve superior results: curve fitting-based for values, and bloom filter-based for indices. DeepReduce is orthogonal to existing gradient sparsifiers and can be applied in conjunction with them, transparently to the end-user, to significantly lower the communication overhead. As proof of concept, we implement our approach on TensorFlow and PyTorch. Our experiments with large real models demonstrate that DeepReduce transmits 320% less data than existing sparsifiers, without affecting accuracy. Code is available at https://github.com/hangxu0304/DeepReduce.
****
Provably Efficient Causal Reinforcement Learning with Confounded Observational Data
Lingxiao Wang, Zhuoran Yang, Zhaoran Wang
Empowered by neural networks, deep reinforcement learning (DRL) achieves tremendous empirical success. However, DRL requires a large dataset by interacting with the environment, which is unrealistic in critical scenarios such as autonomous driving and personalized medicine. In this paper, we study how to incorporate the dataset collected in the offline setting to improve the sample efficiency in the online setting. To incorporate the observational data, we face two challenges. (a) The behavior policy that generates the observational data may depend on unobserved random variables (confounders), which affect the received rewards and t

ransition dynamics. (b) Exploration in the online setting requires quantifying t
he uncertainty given both the observational and interventional data. To tackle s
uch challenges, we propose the deconfounded optimistic value iteration (DOVI) al
gorithm, which incorporates the confounded observational data in a provably effi
cient manner. DOVI explicitly adjusts for the confounding bias in the observatio
nal data, where the confounders are partially observed or unobserved. In both ca
ses, such adjustments allow us to construct the bonus based on a notion of infor
mation gain, which takes into account the amount of information acquired from th
e offline setting. In particular, we prove that the regret of DOVI is smaller th
an the optimal regret achievable in the pure online setting when the confounded
observational data are informative upon the adjustments.
****

Predicting Deep Neural Network Generalization with Perturbation Response Curves
Yair Schiff, Brian Quanz, Payel Das, Pin-Yu Chen
The field of Deep Learning is rich with empirical evidence of human-like perform
ance on a variety of prediction tasks. However, despite these successes, the rec
ent Predicting Generalization in Deep Learning (PGDL) NeurIPS 2020 competition s
uggests that there is a need for more robust and efficient measures of network g
eneralization. In this work, we propose a new framework for evaluating the gener
alization capabilities of trained networks. We use perturbation response (PR) cu
rves that capture the accuracy change of a given network as a function of varyin
g levels of training sample perturbation. From these PR curves, we derive novel
statistics that capture generalization capability. Specifically, we introduce tw
o new measures for accurately predicting generalization gaps: the Gi-score and P
al-score, which are inspired by the Gini coefficient and Palma ratio (measures o
f income inequality), that accurately predict generalization gaps. Using our fra
mework applied to intra and inter-class sample mixup, we attain better predictiv
e scores than the current state-of-the-art measures on a majority of tasks in th
e PGDL competition. In addition, we show that our framework and the proposed sta
tistics can be used to capture to what extent a trained network is invariant to
a given parametric input transformation, such as rotation or translation. Theref
ore, these generalization gap prediction statistics also provide a useful means
for selecting optimal network architectures and hyperparameters that are invaria
nt to a certain perturbation.
****

Exploiting Domain-Specific Features to Enhance Domain Generalization
Manh-Ha Bui, Toan Tran, Anh Tran, Dinh Phung
Domain Generalization (DG) aims to train a model, from multiple observed source
domains, in order to perform well on unseen target domains. To obtain the genera
lization capability, prior DG approaches have focused on extracting domain-invar
iant information across sources to generalize on target domains, while useful do
main-specific information which strongly correlates with labels in individual do
mains and the generalization to target domains is usually ignored. In this paper
, we propose meta-Domain Specific-Domain Invariant (mDSDI) - a novel theoretical
ly sound framework that extends beyond the invariance view to further capture th
e usefulness of domain-specific information. Our key insight is to disentangle f
eatures in the latent space while jointly learning both domain-invariant and dom
ain-specific features in a unified framework. The domain-specific representation
 is optimized through the meta-learning framework to adapt from source domains,
targeting a robust generalization on unseen domains. We empirically show that mD
SDI provides competitive results with state-of-the-art techniques in DG. A furth
er ablation study with our generated dataset, Background-Colored-MNIST, confirms
 the hypothesis that domain-specific is essential, leading to better results whe
n compared with only using domain-invariant.
****

Optimal Order Simple Regret for Gaussian Process Bandits
Sattar Vakili, Nacime Bouziani, Sepehr Jalali, Alberto Bernacchia, Da-shan Shiu
Requests for name changes in the electronic proceedings will be accepted with no
 questions asked. However name changes may cause bibliographic tracking issues.
  Authors are asked to consider this carefully and discuss it with their co-auth

ors prior to requesting a name change in the electronic proceedings.
****
Generalization Guarantee of SGD for Pairwise Learning
Yunwen Lei, Mingrui Liu, Yiming Ying
Recently, there is a growing interest in studying pairwise learning since it inc ludes many important machine learning tasks as specific examples, e.g., metric l earning, AUC maximization and ranking. While stochastic gradient descent (SGD) i s an efficient method, there is a lacking study on its generalization behavior f or pairwise learning. In this paper, we present a systematic study on the genera lization analysis of SGD for pairwise learning to understand the balance between generalization and optimization. We develop a novel high-probability generaliza tion bound for uniformly-stable algorithms to incorporate the variance informati on for better generalization, based on which we establish the first nonsmooth le arning algorithm to achieve almost optimal high-probability and dimension-indepe ndent generalization bounds in linear time. We consider both convex and nonconve x pairwise learning problems. Our stability analysis for convex problems shows h ow the interpolation can help generalization. We establish a uniform convergence of gradients, and apply it to derive the first generalization bounds on populat ion gradients for nonconvex problems. Finally, we develop better generalization bounds for gradient-dominated problems.
****
Supercharging Imbalanced Data Learning With Energy-based Contrastive Representat ion Transfer
Junya Chen, Zidi Xiu, Benjamin Goldstein, Ricardo Henao, Lawrence Carin, Chenyan g Tao
Dealing with severe class imbalance poses a major challenge for many real-world applications, especially when the accurate classification and generalization of minority classes are of primary interest.In computer vision and NLP, learning fr om datasets with long-tail behavior is a recurring theme, especially for natural ly occurring labels. Existing solutions mostly appeal to sampling or weighting a djustments to alleviate the extreme imbalance, or impose inductive bias to prior itize generalizable associations. Here we take a novel perspective to promote sa mple efficiency and model generalization based on the invariance principles of c ausality. Our contribution posits a meta-distributional scenario, where the caus al generating mechanism for label-conditional features is invariant across diffe rent labels. Such causal assumption enables efficient knowledge transfer from th e dominant classes to their under-represented counterparts, even if their featur e distributions show apparent disparities. This allows us to leverage a causal d ata augmentation procedure to enlarge the representation of minority classes. Ou r development is orthogonal to the existing imbalanced data learning techniques thus can be seamlessly integrated. The proposed approach is validated on an exte nsive set of synthetic and real-world tasks against state-of-the-art solutions.
****
Heavy Ball Momentum for Conditional Gradient
Bingcong Li, Alireza Sadeghi, Georgios Giannakis
Conditional gradient, aka Frank Wolfe (FW) algorithms, have well-documented meri ts in machine learning and signal processing applications. Unlike projection-bas ed methods, momentum cannot improve the convergence rate of FW, in general. This limitation motivates the present work, which deals with heavy ball momentum, an d its impact to FW. Specifically, it is established that heavy ball offers a uni fying perspective on the primal-dual (PD) convergence, and enjoys a tighter \tex tit{per iteration} PD error rate, for multiple choices of step sizes, where PD e rror can serve as the stopping criterion in practice. In addition, it is asserte d that restart, a scheme typically employed jointly with Nesterov's momentum, ca n further tighten this PD error bound. Numerical results demonstrate the usefuln ess of heavy ball momentum in FW iterations.
****
PARP: Prune, Adjust and Re-Prune for Self-Supervised Speech Recognition
Cheng-I Jeff Lai, Yang Zhang, Alexander H. Liu, Shiyu Chang, Yi-Lun Liao, Yung-S ung Chuang, Kaizhi Qian, Sameer Khurana, David Cox, Jim Glass

Self-supervised speech representation learning (speech SSL) has demonstrated the benefit of scale in learning rich representations for Automatic Speech Recognition (ASR) with limited paired data, such as wav2vec 2.0. We investigate the existence of sparse subnetworks in pre-trained speech SSL models that achieve even better low-resource ASR results. However, directly applying widely adopted pruning methods such as the Lottery Ticket Hypothesis (LTH) is suboptimal in the computational cost needed. Moreover, we show that the discovered subnetworks yield minimal performance gain compared to the original dense network.We present Prune-Adjust-Re-Prune (PARP), which discovers and finetunes subnetworks for much better performance, while only requiring a single downstream ASR finetuning run. PARP is inspired by our surprising observation that subnetworks pruned for pre-training tasks need merely a slight adjustment to achieve a sizeable performance boost in downstream ASR tasks. Extensive experiments on low-resource ASR verify (1) sparse subnetworks exist in mono-lingual/multi-lingual pre-trained speech SSL, and (2) the computational advantage and performance gain of PARP over baseline pruning methods.In particular, on the 10min Librispeech split without LM decoding, PARP discovers subnetworks from wav2vec 2.0 with an absolute 10.9%/12.6% WER decrease compared to the full model. We further demonstrate the effectiveness of PARP via: cross-lingual pruning without any phone recognition degradation, the discovery of a multi-lingual subnetwork for 10 spoken languages in 1 finetuning run, and its applicability to pre-trained BERT/XLNet for natural language tasks1.
****
Robust Learning of Optimal Auctions
Wenshuo Guo, Michael Jordan, Emmanouil Zampetakis
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Disrupting Deep Uncertainty Estimation Without Harming Accuracy
Ido Galil, Ran El-Yaniv
Deep neural networks (DNNs) have proven to be powerful predictors and are widely used for various tasks. Credible uncertainty estimation of their predictions, however, is crucial for their deployment in many risk-sensitive applications. In this paper we present a novel and simple attack, which unlike adversarial attacks, does not cause incorrect predictions but instead cripples the network's capacity for uncertainty estimation. The result is that after the attack, the DNN is more confident of its incorrect predictions than about its correct ones without having its accuracy reduced. We present two versions of the attack. The first scenario focuses on a black-box regime (where the attacker has no knowledge of the target network) and the second scenario attacks a white-box setting. The proposed attack is only required to be of minuscule magnitude for its perturbations to cause severe uncertainty estimation damage, with larger magnitudes resulting in completely unusable uncertainty estimations.We demonstrate successful attacks on three of the most popular uncertainty estimation methods: the vanilla softmax score, Deep Ensembles and MC-Dropout. Additionally, we show an attack on SelectiveNet, the selective classification architecture. We test the proposed attack on several contemporary architectures such as MobileNetV2 and EfficientNetB0, all trained to classify ImageNet.
****
SOFT: Softmax-free Transformer with Linear Complexity
Jiachen Lu, Jinghan Yao, Junge Zhang, Xiatian Zhu, Hang Xu, Weiguo Gao, Chunjing XU, Tao Xiang, Li Zhang
Vision transformers (ViTs) have pushed the state-of-the-art for various visual recognition tasks by patch-wise image tokenization followed by self-attention. However, the employment of self-attention modules results in a quadratic complexity in both computation and memory usage. Various attempts on approximating the self-attention computation with linear complexity have been made in Natural Language Processing. However, an in-depth analysis in this work shows that they are either theoretically flawed or empirically ineffective for visual recognition. We

further identify that their limitations are rooted in keeping the softmax self-a ttention during approximations. Specifically, conventional self-attention is co mputed by normalizing the scaled dot-product between token feature vectors. Keep ing this softmax operation challenges any subsequent linearization efforts. Base d on this insight, for the first time, a softmax-free transformer or SOFT is pr oposed. To remove softmax in self-attention, Gaussian kernel function is used t o replace the dot-product similarity without further normalization. This enables a full self-attention matrix to be approximated via a low-rank matrix decompos ition. The robustness of the approximation is achieved by calculating its Moore-Penrose inverse using a Newton-Raphson method. Extensive experiments on ImageN et show that our SOFT significantly improves the computational efficiency of exi sting ViT variants. Crucially, with a linear complexity, much longer token seque nces are permitted in SOFT, resulting in superior trade-off between accuracy and complexity.
****
Task-Adaptive Neural Network Search with Meta-Contrastive Learning
Wonyong Jeong, Hayeon Lee, Geon Park, Eunyoung Hyung, Jinheon Baek, Sung Ju Hwan g
Most conventional Neural Architecture Search (NAS) approaches are limited in tha t they only generate architectures without searching for the optimal parameters. While some NAS methods handle this issue by utilizing a supernet trained on a l arge-scale dataset such as ImageNet, they may be suboptimal if the target tasks are highly dissimilar from the dataset the supernet is trained on. To address su ch limitations, we introduce a novel problem of Neural Network Search (NNS), who se goal is to search for the optimal pretrained network for a novel dataset and constraints (e.g. number of parameters), from a model zoo. Then, we propose a no vel framework to tackle the problem, namely Task-Adaptive Neural Network Search (TANS). Given a model-zoo that consists of network pretrained on diverse dataset s, we use a novel amortized meta-learning framework to learn a cross-modal laten t space with contrastive loss, to maximize the similarity between a dataset and a high-performing network on it, and minimize the similarity between irrelevant dataset-network pairs. We validate the effectiveness and efficiency of our metho d on ten real-world datasets, against existing NAS/AutoML baselines. The results show that our method instantly retrieves networks that outperform models obtain ed with the baselines with significantly fewer training steps to reach the targe t performance, thus minimizing the total cost of obtaining a task-optimal networ k. Our code and the model-zoo are available at https://anonymous.4open.science/r /TANS-33D6
****
Neural Flows: Efficient Alternative to Neural ODEs
Marin Biloš, Johanna Sommer, Syama Sundar Rangapuram, Tim Januschowski, Stephan Günnemann
Neural ordinary differential equations describe how values change in time. This is the reason why they gained importance in modeling sequential data, especially when the observations are made at irregular intervals. In this paper we propose an alternative by directly modeling the solution curves - the flow of an ODE - with a neural network. This immediately eliminates the need for expensive numeri cal solvers while still maintaining the modeling capability of neural ODEs. We p ropose several flow architectures suitable for different applications by establi shing precise conditions on when a function defines a valid flow. Apart from com putational efficiency, we also provide empirical evidence of favorable generaliz ation performance via applications in time series modeling, forecasting, and den sity estimation.
****
Multi-Objective Meta Learning
Feiyang YE, Baijiong Lin, Zhixiong Yue, Pengxin Guo, Qiao Xiao, Yu Zhang
Meta learning with multiple objectives has been attracted much attention recentl y since many applications need to consider multiple factors when designing learn ing models. Existing gradient-based works on meta learning with multiple objecti ves mainly combine multiple objectives into a single objective in a weighted sum

manner. This simple strategy usually works but it requires to tune the weights associated with all the objectives, which could be time consuming. Different from those works, in this paper, we propose a gradient-based Multi-Objective Meta Learning (MOML) framework without manually tuning weights. Specifically, MOML formulates the objective function of meta learning with multiple objectives as a Multi-Objective Bi-Level optimization Problem (MOBLP) where the upper-level subproblem is to solve several possibly conflicting objectives for the meta learner. To solve the MOBLP, we devise the first gradient-based optimization algorithm by alternatively solving the lower-level and upper-level subproblems via the gradient descent method and the gradient-based multi-objective optimization method, respectively. Theoretically, we prove the convergence properties of the proposed gradient-based optimization algorithm. Empirically, we show the effectiveness of the proposed MOML framework in several meta learning problems, including few-shot learning, domain adaptation, multi-task learning, and neural architecture search. The source code of MOML is available at https://github.com/Baijiong-Lin/MOML.
****

A self consistent theory of Gaussian Processes captures feature learning effects in finite CNNs

Gadi Naveh, Zohar Ringel

Deep neural networks (DNNs) in the infinite width/channel limit have received much attention recently, as they provide a clear analytical window to deep learning via mappings to Gaussian Processes (GPs). Despite its theoretical appeal, this viewpoint lacks a crucial ingredient of deep learning in finite DNNs, laying at the heart of their success --- \textit{feature learning}. Here we consider DNNs trained with noisy gradient descent on a large training set and derive a self-consistent Gaussian Process theory accounting for \textit{strong} finite-DNN and feature learning effects. Applying this to a toy model of a two-layer linear convolutional neural network (CNN) shows good agreement with experiments. We further identify, both analytically and numerically, a sharp transition between a feature learning regime and a lazy learning regime in this model. Strong finite-DNN effects are also derived for a non-linear two-layer fully connected network. We have numerical evidence demonstrating that the assumptions required for our theory hold true in more realistic settings (Myrtle5 CNN trained on CIFAR-10).Our self-consistent theory provides a rich and versatile analytical framework for studying strong finite-DNN effects, most notably - feature learning.
****

Mini-Batch Consistent Slot Set Encoder for Scalable Set Encoding

Andreis Bruno, Jeffrey Willette, Juho Lee, Sung Ju Hwang

Most existing set encoding algorithms operate under the implicit assumption that all the set elements are accessible, and that there are ample computational and memory resources to load the set into memory during training and inference. However, both assumptions fail when the set is excessively large such that it is impossible to load all set elements into memory, or when data arrives in a stream. To tackle such practical challenges in large-scale set encoding, the general set-function constraints of permutation invariance and equivariance are not sufficient. We introduce a new property termed Mini-Batch Consistency (MBC) that is required for large scale mini-batch set encoding. Additionally, we present a scalable and efficient attention-based set encoding mechanism that is amenable to mini-batch processing of sets, and capable of updating set representations as data arrives. The proposed method adheres to the required symmetries of invariance and equivariance as well as maintaining MBC for any partition of the input set. We perform extensive experiments and show that our method is computationally efficient and results in rich set encoding representations for set-structured data.
****

Efficient and Local Parallel Random Walks

Michael Kapralov, Silvio Lattanzi, Navid Nouri, Jakab Tardos

Random walks are a fundamental primitive used in many machine learning algorithms with several applications in clustering and semi-supervised learning. Despite their relevance, the first efficient parallel algorithm to compute random walks

has been introduced very recently (■■cki et al.). Unfortunately their method has
 a fundamental shortcoming: their algorithm is non-local in that it heavily reli
es on computing random walks out of all nodes in the input graph, even though in
 many practical applications one is interested in computing random walks only fr
om a small subset of nodes in the graph. In this paper, we present a new algorit
hm that overcomes this limitation by building random walks efficiently and loca
lly at the same time. We show that our technique is both memory and round effici
ent, and in particular yields an efficient parallel local clustering algorithm.
Finally, we complement our theoretical analysis with experimental results showin
g that our algorithm is significantly more scalable than previous approaches.
****
Amortized Variational Inference for Simple Hierarchical Models
Abhinav Agrawal, Justin Domke
It is difficult to use subsampling with variational inference in hierarchical mo
dels since the number of local latent variables scales with the dataset. Thus, i
nference in hierarchical models remains a challenge at a large scale. It is help
ful to use a variational family with a structure matching the posterior, but opt
imization is still slow due to the huge number of local distributions. Instead,
this paper suggests an amortized approach where shared parameters simultaneously
 represent all local distributions. This approach is similarly accurate as using
 a given joint distribution (e.g., a full-rank Gaussian) but is feasible on data
sets that are several orders of magnitude larger. It is also dramatically faster
 than using a structured variational distribution.
****
Online Matching in Sparse Random Graphs: Non-Asymptotic Performances of Greedy A
lgorithm
Nathan Noiry, Vianney Perchet, Flore Sentenac
Motivated by sequential budgeted allocation problems, we investigate online mat
ching problems where connections between vertices are not i.i.d., but they have
fixed degree distributions -- the so-called configuration model. We estimate the
 competitive ratio of the simplest algorithm, GREEDY, by approximating some rele
vant stochastic discrete processes by their continuous counterparts, that are so
lutions of an explicit system of partial differential equations. This technique
gives precise bounds on the estimation errors, with arbitrarily high probabili
ty as the problem size increases. In particular, it allows the formal comparison
 between different configuration models. We also prove that, quite surprisingly,
  GREEDY can have better performance guarantees than RANKING, another celebrate
d algorithm for online matching that usually outperforms the former.
****
End-to-end reconstruction meets data-driven regularization for inverse problems
Subhadip Mukherjee, Marcello Carioni, Ozan Öktem, Carola-Bibiane Schönlieb
We propose a new approach for learning end-to-end reconstruction operators based
 on unpaired training data for ill-posed inverse problems. The proposed method c
ombines the classical variational framework with iterative unrolling and essenti
ally seeks to minimize a weighted combination of the expected distortion in the
measurement space and the Wasserstein-1 distance between the distributions of th
e reconstruction and the ground-truth. More specifically, the regularizer in the
 variational setting is parametrized by a deep neural network and learned simult
aneously with the unrolled reconstruction operator. The variational problem is t
hen initialized with the output of the reconstruction network and solved iterati
vely till convergence. Notably, it takes significantly fewer iterations to conve
rge as compared to variational methods, thanks to the excellent initialization o
btained via the unrolled operator. The resulting approach combines the computati
onal efficiency of end-to-end unrolled reconstruction with the well-posedness an
d noise-stability guarantees of the variational setting. Moreover, we demonstrat
e with the example of image reconstruction in X-ray computed tomography (CT) tha
t our approach outperforms state-of-the-art unsupervised methods and that it out
performs or is at least on par with state-of-the-art supervised data-driven reco
nstruction approaches.
****

An online passive-aggressive algorithm for difference-of-squares classification
Lawrence Saul
We investigate a low-rank model of quadratic classification inspired by previous work on factorization machines, polynomial networks, and capsule-based architectures for visual object recognition. The model is parameterized by a pair of affine transformations, and it classifies examples by comparing the magnitudes of vectors that these transformations produce. The model is also over-parameterized in the sense that different pairs of affine transformations can describe classifiers with the same decision boundary and confidence scores. We show that such pairs arise from discrete and continuous symmetries of the model's parameter space: in particular, the latter define symmetry groups of rotations and Lorentz transformations, and we use these group structures to devise appropriately invariant procedures for model alignment and averaging. We also leverage the form of the model's decision boundary to derive simple margin-based updates for online learning. Here we explore a strategy of passive-aggressive learning: for each example, we compute the minimum change in parameters that is required to predict its correct label with high confidence. We derive these updates by solving a quadratically constrained quadratic program (QCQP); interestingly, this QCQP is nonconvex but tractable, and it can be solved efficiently by elementary methods. We highlight the conceptual and practical contributions of this approach. Conceptually, we show that it extends the paradigm of passive-aggressive learning to a larger family of nonlinear models for classification. Practically, we show that these models perform well on large-scale problems in online learning.
****
Finite-Sample Analysis of Off-Policy TD-Learning via Generalized Bellman Operators
Zaiwei Chen, Siva Theja Maguluri, Sanjay Shakkottai, Karthikeyan Shanmugam
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 Bi-Level Framework for Learning to Solve Combinatorial Optimization on Graphs
Runzhong Wang, Zhigang Hua, Gan Liu, Jiayi Zhang, Junchi Yan, Feng Qi, Shuang Yang, Jun Zhou, Xiaokang Yang
Combinatorial Optimization (CO) has been a long-standing challenging research topic featured by its NP-hard nature. Traditionally such problems are approximately solved with heuristic algorithms which are usually fast but may sacrifice the solution quality. Currently, machine learning for combinatorial optimization (MLCO) has become a trending research topic, but most existing MLCO methods treat CO as a single-level optimization by directly learning the end-to-end solutions, which are hard to scale up and mostly limited by the capacity of ML models given the high complexity of CO. In this paper, we propose a hybrid approach to combine the best of the two worlds, in which a bi-level framework is developed with an upper-level learning method to optimize the graph (e.g. add, delete or modify edges in a graph), fused with a lower-level heuristic algorithm solving on the optimized graph. Such a bi-level approach simplifies the learning on the original hard CO and can effectively mitigate the demand for model capacity. The experiments and results on several popular CO problems like Directed Acyclic Graph scheduling, Graph Edit Distance and Hamiltonian Cycle Problem show its effectiveness over manually designed heuristics and single-level learning methods.
****
Improved Learning Rates of a Functional Lasso-type SVM with Sparse Multi-Kernel Representation
shaogao lv, Junhui Wang, Jiankun Liu, Yong Liu
In this paper, we provide theoretical results of estimation bounds and excess risk upper bounds for support vector machine (SVM) with sparse multi-kernel representation. These convergence rates for multi-kernel SVM are established by analyzing a Lasso-type regularized learning scheme within composite multi-kernel spaces. It is shown that the oracle rates of convergence of classifiers depend on

the complexity of multi-kernels, the sparsity, a Bernstein condition and the sample size, which significantly improves on previous results even for the additive or linear cases. In summary, this paper not only provides unified theoretical results for multi-kernel SVMs, but also enriches the literature on high-dimensional nonparametric classification.
****
When does Contrastive Learning Preserve Adversarial Robustness from Pretraining to Finetuning?
Lijie Fan, Sijia Liu, Pin-Yu Chen, Gaoyuan Zhang, Chuang Gan
Contrastive learning (CL) can learn generalizable feature representations and achieve state-of-the-art performance of downstream tasks by finetuning a linear classifier on top of it. However, as adversarial robustness becomes vital in image classification, it remains unclear whether or not CL is able to preserve robustness to downstream tasks. The main challenge is that in the self-supervised pretraining + supervised finetuning paradigm, adversarial robustness is easily forgotten due to a learning task mismatch from pretraining to finetuning. We call such challenge 'cross-task robustness transferability'. To address the above problem, in this paper we revisit and advance CL principles through the lens of robustness enhancement. We show that (1) the design of contrastive views matters: High-frequency components of images are beneficial to improving model robustness; (2) Augmenting CL with pseudo-supervision stimulus (e.g., resorting to feature clustering) helps preserve robustness without forgetting. Equipped with our new designs, we propose AdvCL, a novel adversarial contrastive pretraining framework. We show that AdvCL is able to enhance cross-task robustness transferability without loss of model accuracy and finetuning efficiency. With a thorough experimental study, we demonstrate that AdvCL outperforms the state-of-the-art self-supervised robust learning methods across multiple datasets (CIFAR-10, CIFAR-100, and STL-10) and finetuning schemes (linear evaluation and full model finetuning).
****
Learning Transferable Features for Point Cloud Detection via 3D Contrastive Co-training
Zeng Yihan, Chunwei Wang, Yunbo Wang, Hang Xu, Chaoqiang Ye, Zhen Yang, Chao Ma
Most existing point cloud detection models require large-scale, densely annotated datasets. They typically underperform in domain adaptation settings, due to geometry shifts caused by different physical environments or LiDAR sensor configurations. Therefore, it is challenging but valuable to learn transferable features between a labeled source domain and a novel target domain, without any access to target labels. To tackle this problem, we introduce the framework of 3D Contrastive Co-training (3D-CoCo) with two technical contributions. First, 3D-CoCo is inspired by our observation that the bird-eye-view (BEV) features are more transferable than low-level geometry features. We thus propose a new co-training architecture that includes separate 3D encoders with domain-specific parameters, as well as a BEV transformation module for learning domain-invariant features. Second, 3D-CoCo extends the approach of contrastive instance alignment to point cloud detection, whose performance was largely hindered by the mismatch between the fictitious distribution of BEV features, induced by pseudo-labels, and the true distribution. The mismatch is greatly reduced by 3D-CoCo with transformed point clouds, which are carefully designed by considering specific geometry priors. We construct new domain adaptation benchmarks using three large-scale 3D datasets. Experimental results show that our proposed 3D-CoCo effectively closes the domain gap and outperforms the state-of-the-art methods by large margins.
****
SILG: The Multi-domain Symbolic Interactive Language Grounding Benchmark
Victor Zhong, Austin W. Hanjie, Sida Wang, Karthik Narasimhan, Luke Zettlemoyer
Existing work in language grounding typically study single environments. How do we build unified models that apply across multiple environments? We propose the multi-environment Symbolic Interactive Language Grounding benchmark (SILG), which unifies a collection of diverse grounded language learning environments under a common interface. SILG consists of grid-world environments that require genera

lization to new dynamics, entities, and partially observed worlds (RTFM, Messeng er, NetHack), as well as symbolic counterparts of visual worlds that re- quire i nterpreting rich natural language with respect to complex scenes (ALFWorld, Touc hdown). Together, these environments provide diverse grounding challenges in ric hness of observation space, action space, language specification, and plan com- plexity. In addition, we propose the first shared model architecture for RL on t hese environments, and evaluate recent advances such as egocentric local convolu tion, recurrent state-tracking, entity-centric attention, and pretrained LM usin g SILG. Our shared architecture achieves comparable performance to environment-s pecific architectures. Moreover, we find that many recent modelling advances do not result in significant gains on environments other than the one they were des igned for. This highlights the need for a multi-environment benchmark. Finally, the best models significantly underperform humans on SILG, which suggests ample room for future work. We hope SILG enables the community to quickly identify new methodolo- gies for language grounding that generalize to a diverse set of envi ronments and their associated challenges.
****
A Surrogate Objective Framework for Prediction+Programming with Soft Constraints
Kai Yan, Jie Yan, Chuan Luo, Liting Chen, Qingwei Lin, Dongmei Zhang
Prediction+optimization is a common real-world paradigm where we have to predict problem parameters before solving the optimization problem. However, the criter ia by which the prediction model is trained are often inconsistent with the goal of the downstream optimization problem. Recently, decision-focused prediction approaches, such as SPO+ and direct optimization, have been proposed to fill thi s gap. However, they cannot directly handle the soft constraints with the max o perator required in many real-world objectives. This paper proposes a novel ana lytically differentiable surrogate objective framework for real-world linear and semi-definite negative quadratic programming problems with soft linear and non-negative hard constraints. This framework gives the theoretical bounds on constr aints' multipliers, and derives the closed-form solution with respect to predict ive parameters and thus gradients for any variable in the problem. We evaluate our method in three applications extended with soft constraints: synthetic linea r programming, portfolio optimization, and resource provisioning, demonstrating that our method outperforms traditional two-staged methods and other decision-fo cused approaches
****
Learning to Predict Trustworthiness with Steep Slope Loss
Yan Luo, Yongkang Wong, Mohan S. Kankanhalli, Qi Zhao
Understanding the trustworthiness of a prediction yielded by a classifier is cri tical for the safe and effective use of AI models. Prior efforts have been prove n to be reliable on small-scale datasets. In this work, we study the problem of predicting trustworthiness on real-world large-scale datasets, where the task is more challenging due to high-dimensional features, diverse visual concepts, and a large number of samples. In such a setting, we observe that the trustworthine ss predictors trained with prior-art loss functions, i.e., the cross entropy los s, focal loss, and true class probability confidence loss, are prone to view bot h correct predictions and incorrect predictions to be trustworthy. The reasons a re two-fold. Firstly, correct predictions are generally dominant over incorrect predictions. Secondly, due to the data complexity, it is challenging to differen tiate the incorrect predictions from the correct ones on real-world large-scale datasets. To improve the generalizability of trustworthiness predictors, we prop ose a novel steep slope loss to separate the features w.r.t. correct predictions from the ones w.r.t. incorrect predictions by two slide-like curves that oppose each other. The proposed loss is evaluated with two representative deep learnin g models, i.e., Vision Transformer and ResNet, as trustworthiness predictors. We conduct comprehensive experiments and analyses on ImageNet, which show that the proposed loss effectively improves the generalizability of trustworthiness pred ictors. The code and pre-trained trustworthiness predictors for reproducibility are available at \url{https://github.com/luoyan407/predict_trustworthiness}.
****

On the Periodic Behavior of Neural Network Training with Batch Normalization and Weight Decay
Ekaterina Lobacheva, Maxim Kodryan, Nadezhda Chirkova, Andrey Malinin, Dmitry P. Vetrov
Training neural networks with batch normalization and weight decay has become a common practice in recent years. In this work, we show that their combined use may result in a surprising periodic behavior of optimization dynamics: the training process regularly exhibits destabilizations that, however, do not lead to complete divergence but cause a new period of training. We rigorously investigate the mechanism underlying the discovered periodic behavior from both empirical and theoretical points of view and analyze the conditions in which it occurs in practice. We also demonstrate that periodic behavior can be regarded as a generalization of two previously opposing perspectives on training with batch normalization and weight decay, namely the equilibrium presumption and the instability presumption.
****
NeRV: Neural Representations for Videos
Hao Chen, Bo He, Hanyu Wang, Yixuan Ren, Ser Nam Lim, Abhinav Shrivastava
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Surrogate Regret Bounds for Polyhedral Losses
Rafael Frongillo, Bo Waggoner
Surrogate risk minimization is an ubiquitous paradigm in supervised machine learning, wherein a target problem is solved by minimizing a surrogate loss on a dataset. Surrogate regret bounds, also called excess risk bounds, are a common tool to prove generalization rates for surrogate risk minimization. While surrogate regret bounds have been developed for certain classes of loss functions, such as proper losses, general results are relatively sparse. We provide two general results. The first gives a linear surrogate regret bound for any polyhedral (piecewise-linear and convex) surrogate, meaning that surrogate generalization rates translate directly to target rates. The second shows that for sufficiently non-polyhedral surrogates, the regret bound is a square root, meaning fast surrogate generalization rates translate to slow rates for the target. Together, these results suggest polyhedral surrogates are optimal in many cases.
****
Last iterate convergence of SGD for Least-Squares in the Interpolation regime.
Aditya Vardhan Varre, Loucas Pillaud-Vivien, 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.
****
Generative vs. Discriminative: Rethinking The Meta-Continual Learning
Mohammadamin Banayeeanzade, Rasoul Mirzaiezadeh, Hosein Hasani, Mahdieh Soleymani
Deep neural networks have achieved human-level capabilities in various learning tasks. However, they generally lose performance in more realistic scenarios like learning in a continual manner. In contrast, humans can incorporate their prior knowledge to learn new concepts efficiently without forgetting older ones. In this work, we leverage meta-learning to encourage the model to learn how to learn continually. Inspired by human concept learning, we develop a generative classifier that efficiently uses data-driven experience to learn new concepts even from few samples while being immune to forgetting. Along with cognitive and theoretical insights, extensive experiments on standard benchmarks demonstrate the effectiveness of the proposed method. The ability to remember all previous concepts, with negligible computational and structural overheads, suggests that generative models provide a natural way for alleviating catastrophic forgetting, which is

a major drawback of discriminative models.
****
Model, sample, and epoch-wise descents: exact solution of gradient flow in the random feature model
Antoine Bodin, Nicolas Macris
Recent evidence has shown the existence of a so-called double-descent and even triple-descent behavior for the generalization error of deep-learning models. This important phenomenon commonly appears in implemented neural network architectures, and also seems to emerge in epoch-wise curves during the training process. A recent line of research has highlighted that random matrix tools can be used to obtain precise analytical asymptotics of the generalization (and training) errors of the random feature model. In this contribution, we analyze the whole temporal behavior of the generalization and training errors under gradient flow for the random feature model. We show that in the asymptotic limit of large system size the full time-evolution path of both errors can be calculated analytically. This allows us to observe how the double and triple descents develop over time, if and when early stopping is an option, and also observe time-wise descent structures. Our techniques are based on Cauchy complex integral representations of the errors together with recent random matrix methods based on linear pencils.
****
Rethinking Graph Transformers with Spectral Attention
Devin Kreuzer, Dominique Beaini, Will Hamilton, Vincent Létourneau, Prudencio Tossou
In recent years, the Transformer architecture has proven to be very successful in sequence processing, but its application to other data structures, such as graphs, has remained limited due to the difficulty of properly defining positions. Here, we present the \textit{Spectral Attention Network} (SAN), which uses a learned positional encoding (LPE) that can take advantage of the full Laplacian spectrum to learn the position of each node in a given graph.This LPE is then added to the node features of the graph and passed to a fully-connected Transformer.By leveraging the full spectrum of the Laplacian, our model is theoretically powerful in distinguishing graphs, and can better detect similar sub-structures from their resonance.Further, by fully connecting the graph, the Transformer does not suffer from over-squashing, an information bottleneck of most GNNs, and enables better modeling of physical phenomenons such as heat transfer and electric interaction.When tested empirically on a set of 4 standard datasets, our model performs on par or better than state-of-the-art GNNs, and outperforms any attention-based model by a wide margin, becoming the first fully-connected architecture to perform well on graph benchmarks.
****
Perceptual Score: What Data Modalities Does Your Model Perceive?
Itai Gat, Idan Schwartz, Alex Schwing
Machine learning advances in the last decade have relied significantly on large-scale datasets that continue to grow in size. Increasingly, those datasets also contain different data modalities. However, large multi-modal datasets are hard to annotate, and annotations may contain biases that we are often unaware of. Deep-net-based classifiers, in turn, are prone to exploit those biases and to find shortcuts. To study and quantify this concern, we introduce the perceptual score, a metric that assesses the degree to which a model relies on the different subsets of the input features, i.e., modalities. Using the perceptual score, we find a surprisingly consistent trend across four popular datasets: recent, more accurate state-of-the-art multi-modal models for visual question-answering or visual dialog tend to perceive the visual data less than their predecessors. This is concerning as answers are hence increasingly inferred from textual cues only. Using the perceptual score also helps to analyze model biases by decomposing the score into data subset contributions. We hope to spur a discussion on the perceptiveness of multi-modal models and also hope to encourage the community working on multi-modal classifiers to start quantifying perceptiveness via the proposed perceptual score.
****

PiRank: Scalable Learning To Rank via Differentiable Sorting
Robin Swezey, Aditya Grover, Bruno Charron, Stefano Ermon
A key challenge with machine learning approaches for ranking is the gap between
the performance metrics of interest and the surrogate loss functions that can be
 optimized with gradient-based methods. This gap arises because ranking metrics
typically involve a sorting operation which is not differentiable w.r.t. the mod
el parameters. Prior works have proposed surrogates that are loosely related to
ranking metrics or simple smoothed versions thereof, and often fail to scale to
real-world applications. We propose PiRank, a new class of differentiable surrog
ates for ranking, which employ a continuous, temperature-controlled relaxation t
o the sorting operator based on NeuralSort [1]. We show that PiRank exactly reco
vers the desired metrics in the limit of zero temperature and further propose a
divide-and-conquer extension that scales favorably to large list sizes, both in
theory and practice. Empirically, we demonstrate the role of larger list sizes d
uring training and show that PiRank significantly improves over comparable appro
aches on publicly available Internet-scale learning-to-rank benchmarks.
****
Deceive D: Adaptive Pseudo Augmentation for GAN Training with Limited Data
Liming Jiang, Bo Dai, Wayne Wu, Chen Change Loy
Generative adversarial networks (GANs) typically require ample data for training
 in order to synthesize high-fidelity images. Recent studies have shown that tra
ining GANs with limited data remains formidable due to discriminator overfitting
, the underlying cause that impedes the generator's convergence. This paper intr
oduces a novel strategy called Adaptive Pseudo Augmentation (APA) to encourage h
ealthy competition between the generator and the discriminator. As an alternativ
e method to existing approaches that rely on standard data augmentations or mode
l regularization, APA alleviates overfitting by employing the generator itself t
o augment the real data distribution with generated images, which deceives the d
iscriminator adaptively. Extensive experiments demonstrate the effectiveness of
APA in improving synthesis quality in the low-data regime. We provide a theoreti
cal analysis to examine the convergence and rationality of our new training stra
tegy. APA is simple and effective. It can be added seamlessly to powerful contem
porary GANs, such as StyleGAN2, with negligible computational cost. Code: https:
//github.com/EndlessSora/DeceiveD.
****
CoFrNets: Interpretable Neural Architecture Inspired by Continued Fractions
Isha Puri, Amit Dhurandhar, Tejaswini Pedapati, Karthikeyan Shanmugam, Dennis We
i, Kush R. Varshney
In recent years there has been a considerable amount of research on local post h
oc explanations for neural networks. However, work on building interpretable neu
ral architectures has been relatively sparse. In this paper, we present a novel
neural architecture, CoFrNet, inspired by the form of continued fractions which
are known to have many attractive properties in number theory, such as fast conv
ergence of approximations to real numbers. We show that CoFrNets can be efficien
tly trained as well as interpreted leveraging their particular functional form.
Moreover, we prove that such architectures are universal approximators based on
a proof strategy that is different than the typical strategy used to prove unive
rsal approximation results for neural networks based on infinite width (or depth
), which is likely to be of independent interest. We experiment on nonlinear syn
thetic functions and are able to accurately model as well as estimate feature at
tributions and even higher order terms in some cases, which is a testament to th
e representational power as well as interpretability of such architectures. To f
urther showcase the power of CoFrNets, we experiment on seven real datasets span
ning tabular, text and image modalities, and show that they are either comparabl
e or significantly better than other interpretable models and multilayer percept
rons, sometimes approaching the accuracies of state-of-the-art models.
****
Iterative Teaching by Label Synthesis
Weiyang Liu, Zhen Liu, Hanchen Wang, Liam Paull, Bernhard Schölkopf, Adrian Well
er

In this paper, we consider the problem of iterative machine teaching, where a teacher provides examples sequentially based on the current iterative learner. In contrast to previous methods that have to scan over the entire pool and select teaching examples from it in each iteration, we propose a label synthesis teaching framework where the teacher randomly selects input teaching examples (e.g., images) and then synthesizes suitable outputs (e.g., labels) for them. We show that this framework can avoid costly example selection while still provably achieving exponential teachability. We propose multiple novel teaching algorithms in this framework. Finally, we empirically demonstrate the value of our framework.
****

Variational Diffusion Models

Diederik Kingma, Tim Salimans, Ben Poole, Jonathan Ho

Diffusion-based generative models have demonstrated a capacity for perceptually impressive synthesis, but can they also be great likelihood-based models? We answer this in the affirmative, and introduce a family of diffusion-based generative models that obtain state-of-the-art likelihoods on standard image density estimation benchmarks. Unlike other diffusion-based models, our method allows for efficient optimization of the noise schedule jointly with the rest of the model. We show that the variational lower bound (VLB) simplifies to a remarkably short expression in terms of the signal-to-noise ratio of the diffused data, thereby improving our theoretical understanding of this model class. Using this insight, we prove an equivalence between several models proposed in the literature. In addition, we show that the continuous-time VLB is invariant to the noise schedule, except for the signal-to-noise ratio at its endpoints. This enables us to learn a noise schedule that minimizes the variance of the resulting VLB estimator, leading to faster optimization. Combining these advances with architectural improvements, we obtain state-of-the-art likelihoods on image density estimation benchmarks, outperforming autoregressive models that have dominated these benchmarks for many years, with often significantly faster optimization. In addition, we show how to use the model as part of a bits-back compression scheme, and demonstrate lossless compression rates close to the theoretical optimum.
****

FastCorrect: Fast Error Correction with Edit Alignment for Automatic Speech Recognition

Yichong Leng, Xu Tan, Linchen Zhu, Jin Xu, Renqian Luo, Linquan Liu, Tao Qin, Xiangyang Li, Edward Lin, Tie-Yan Liu

Error correction techniques have been used to refine the output sentences from automatic speech recognition (ASR) models and achieve a lower word error rate (WER) than original ASR outputs. Previous works usually use a sequence-to-sequence model to correct an ASR output sentence autoregressively, which causes large latency and cannot be deployed in online ASR services. A straightforward solution to reduce latency, inspired by non-autoregressive (NAR) neural machine translation, is to use an NAR sequence generation model for ASR error correction, which, however, comes at the cost of significantly increased ASR error rate. In this paper, observing distinctive error patterns and correction operations (i.e., insertion, deletion, and substitution) in ASR, we propose FastCorrect, a novel NAR error correction model based on edit alignment. In training, FastCorrect aligns each source token from an ASR output sentence to the target tokens from the corresponding ground-truth sentence based on the edit distance between the source and target sentences, and extracts the number of target tokens corresponding to each source token during edition/correction, which is then used to train a length predictor and to adjust the source tokens to match the length of the target sentence for parallel generation. In inference, the token number predicted by the length predictor is used to adjust the source tokens for target sequence generation. Experiments on the public AISHELL-1 dataset and an internal industrial-scale ASR dataset show the effectiveness of FastCorrect for ASR error correction: 1) it speeds up the inference by 6-9 times and maintains the accuracy (8-14% WER reduction) compared with the autoregressive correction model; and 2) it outperforms the popular NAR models adopted in neural machine translation and text edition by a large margin.

****
Integrated Latent Heterogeneity and Invariance Learning in Kernel Space
Jiashuo Liu, Zheyuan Hu, Peng Cui, Bo Li, Zheyan Shen
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Hierarchical Reinforcement Learning with Timed Subgoals
Nico Gürtler, Dieter Büchler, Georg Martius
Hierarchical reinforcement learning (HRL) holds great potential for sample-efficient learning on challenging long-horizon tasks. In particular, letting a higher level assign subgoals to a lower level has been shown to enable fast learning on difficult problems. However, such subgoal-based methods have been designed with static reinforcement learning environments in mind and consequently struggle with dynamic elements beyond the immediate control of the agent even though they are ubiquitous in real-world problems. In this paper, we introduce Hierarchical reinforcement learning with Timed Subgoals (HiTS), an HRL algorithm that enables the agent to adapt its timing to a dynamic environment by not only specifying what goal state is to be reached but also when. We discuss how communicating with a lower level in terms of such timed subgoals results in a more stable learning problem for the higher level. Our experiments on a range of standard benchmarks and three new challenging dynamic reinforcement learning environments show that our method is capable of sample-efficient learning where an existing state-of-the-art subgoal-based HRL method fails to learn stable solutions.
****
Fair Scheduling for Time-dependent Resources
Bo Li, Minming Li, Ruilong Zhang
We study a fair resource scheduling problem, where a set of interval jobs are to be allocated to heterogeneous machines controlled by intellectual agents.Each job is associated with release time, deadline, and processing time such that it can be processed if its complete processing period is between its release time and deadline. The machines gain possibly different utilities by processing different jobs, and all jobs assigned to the same machine should be processed without overlap.We consider two widely studied solution concepts, namely, maximin share fairness and envy-freeness.For both criteria, we discuss the extent to which fair allocations exist and present constant approximation algorithms for various settings.
****
SNIPS: Solving Noisy Inverse Problems Stochastically
Bahjat Kawar, Gregory Vaksman, Michael Elad
In this work we introduce a novel stochastic algorithm dubbed SNIPS, which draws samples from the posterior distribution of any linear inverse problem, where the observation is assumed to be contaminated by additive white Gaussian noise. Our solution incorporates ideas from Langevin dynamics and Newton's method, and exploits a pre-trained minimum mean squared error (MMSE) Gaussian denoiser. The proposed approach relies on an intricate derivation of the posterior score function that includes a singular value decomposition (SVD) of the degradation operator, in order to obtain a tractable iterative algorithm for the desired sampling. Due to its stochasticity, the algorithm can produce multiple high perceptual quality samples for the same noisy observation. We demonstrate the abilities of the proposed paradigm for image deblurring, super-resolution, and compressive sensing. We show that the samples produced are sharp, detailed and consistent with the given measurements, and their diversity exposes the inherent uncertainty in the inverse problem being solved.
****
Stateful ODE-Nets using Basis Function Expansions
Alejandro Queiruga, N. Benjamin Erichson, Liam Hodgkinson, Michael W. Mahoney
The recently-introduced class of ordinary differential equation networks (ODE-Nets) establishes a fruitful connection between deep learning and dynamical system

s. In this work, we reconsider formulations of the weights as continuous-in-dept
h functions using linear combinations of basis functions which enables us to lev
erage parameter transformations such as function projections. In turn, this view
 allows us to formulate a novel stateful ODE-Block that handles stateful layers.
 The benefits of this new ODE-Block are twofold: first, it enables incorporating
 meaningful continuous-in-depth batch normalization layers to achieve state-of-t
he-art performance; second, it enables compressing the weights through a change
of basis, without retraining, while maintaining near state-of-the-art performanc
e and reducing both inference time and memory footprint. Performance is demonstr
ated by applying our stateful ODE-Block to (a) image classification tasks using
convolutional units and (b) sentence-tagging tasks using transformer encoder uni
ts.
****
Beyond the Signs: Nonparametric Tensor Completion via Sign Series
Chanwoo Lee, Miaoyan Wang
We consider the problem of tensor estimation from noisy observations with possib
ly missing entries. A nonparametric approach to tensor completion is developed b
ased on a new model which we coin as sign representable tensors. The model repre
sents the signal tensor of interest using a series of structured sign tensors. U
nlike earlier methods, the sign series representation effectively addresses both
 low- and high-rank signals, while encompassing many existing tensor models---in
cluding CP models, Tucker models, single index models, structured tensors with r
epeating entries---as special cases. We provably reduce the tensor estimation pr
oblem to a series of structured classification tasks, and we develop a learning
reduction machinery to empower existing low-rank tensor algorithms for more chal
lenging high-rank estimation. Excess risk bounds, estimation errors, and sample
complexities are established. We demonstrate the outperformance of our approach
over previous methods on two datasets, one on human brain connectivity networks
and the other on topic data mining.
****
Functional Variational Inference based on Stochastic Process Generators
Chao Ma, José Miguel Hernández-Lobato
Bayesian inference in the space of functions has been an important topic for Bay
esian modeling in the past. In this paper, we propose a new solution to this pro
blem called Functional Variational Inference (FVI). In FVI, we minimize a diverg
ence in function space between the variational distribution and the posterior pr
ocess. This is done by using as functional variational family a new class of fle
xible distributions called Stochastic Process Generators (SPGs), which are cleve
rly designed so that the functional ELBO can be estimated efficiently using anal
ytic solutions and mini-batch sampling. FVI can be applied to stochastic process
 priors when random function samples from those priors are available. Our experi
ments show that FVI consistently outperforms weight-space and function space VI
methods on several tasks, which validates the effectiveness of our approach.
****
TTT++: When Does Self-Supervised Test-Time Training Fail or Thrive?
Yuejiang Liu, Parth Kothari, Bastien van Delft, Baptiste Bellot-Gurlet, Taylor M
ordan, Alexandre Alahi
Test-time training (TTT) through self-supervised learning (SSL) is an emerging p
aradigm to tackle distributional shifts. Despite encouraging results, it remains
 unclear when this approach thrives or fails. In this work, we first provide an
in-depth look at its limitations and show that TTT can possibly deteriorate, ins
tead of improving, the test-time performance in the presence of severe distribut
ion shifts. To address this issue, we introduce a test-time feature alignment st
rategy utilizing offline feature summarization and online moment matching, which
 regularizes adaptation without revisiting training data. We further scale this
strategy in the online setting through batch-queue decoupling to enable robust m
oment estimates even with limited batch size. Given aligned feature distribution
s, we then shed light on the strong potential of TTT by theoretically analyzing
its performance post adaptation. This analysis motivates our use of more informa
tive self-supervision in the form of contrastive learning for visual recognition

problems. We empirically demonstrate that our modified version of test-time training, termed TTT++, outperforms state-of-the-art methods by significant margins on several benchmarks. Our result indicates that storing and exploiting extra information, in addition to model parameters, can be a promising direction towards robust test-time adaptation.
****
Double Machine Learning Density Estimation for Local Treatment Effects with Instruments
Yonghan Jung, Jin Tian, Elias Bareinboim
Local treatment effects are a common quantity found throughout the empirical sciences that measure the treatment effect among those who comply with what they are assigned. Most of the literature is focused on estimating the average of such quantity, which is called the ``local average treatment effect (LATE)'' [Imbens and Angrist, 1994]). In this work, we study how to estimate the density of the local treatment effect, which is naturally more informative than its average. Specifically, we develop two families of methods for this task, namely, kernel-smoothing and model-based approaches. The kernel-smoothing-based approach estimates the density through some smooth kernel functions. The model-based approach estimates the density by projecting it onto a finite-dimensional density class. For both approaches, we derive the corresponding double/debiased machine learning-based estimators [Chernozhukov et al., 2018]. We further study the asymptotic convergence rates of the estimators and show that they are robust to the biases in nuisance function estimation. The use of the proposed methods is illustrated through both synthetic and a real dataset called 401(k).
****
Dirichlet Energy Constrained Learning for Deep Graph Neural Networks
Kaixiong Zhou, Xiao Huang, Daochen Zha, Rui Chen, Li Li, Soo-Hyun Choi, Xia Hu
Graph neural networks (GNNs) integrate deep architectures and topological structure modeling in an effective way. However, the performance of existing GNNs would decrease significantly when they stack many layers, because of the over-smoothing issue. Node embeddings tend to converge to similar vectors when GNNs keep recursively aggregating the representations of neighbors. To enable deep GNNs, several methods have been explored recently. But they are developed from either techniques in convolutional neural networks or heuristic strategies. There is no generalizable and theoretical principle to guide the design of deep GNNs. To this end, we analyze the bottleneck of deep GNNs by leveraging the Dirichlet energy of node embeddings, and propose a generalizable principle to guide the training of deep GNNs. Based on it, a novel deep GNN framework -- Energetic Graph Neural Networks (EGNN) is designed. It could provide lower and upper constraints in terms of Dirichlet energy at each layer to avoid over-smoothing. Experimental results demonstrate that EGNN achieves state-of-the-art performance by using deep layers.
****
Accelerating Robotic Reinforcement Learning via Parameterized Action Primitives
Murtaza Dalal, Deepak Pathak, Russ R. Salakhutdinov
Despite the potential of reinforcement learning (RL) for building general-purpose robotic systems, training RL agents to solve robotics tasks still remains challenging due to the difficulty of exploration in purely continuous action spaces. Addressing this problem is an active area of research with the majority of focus on improving RL methods via better optimization or more efficient exploration. An alternate but important component to consider improving is the interface of the RL algorithm with the robot. In this work, we manually specify a library of robot action primitives (RAPS), parameterized with arguments that are learned by an RL policy. These parameterized primitives are expressive, simple to implement, enable efficient exploration and can be transferred across robots, tasks and environments. We perform a thorough empirical study across challenging tasks in three distinct domains with image input and a sparse terminal reward. We find that our simple change to the action interface substantially improves both the learning efficiency and task performance irrespective of the underlying RL algorithm, significantly outperforming prior methods which learn skills from offline ex

pert data.
****

Boosted CVaR Classification
Runtian Zhai, Chen Dan, Arun Suggala, J. Zico Kolter, Pradeep Ravikumar
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

Disentangled Contrastive Learning on Graphs
Haoyang Li, Xin Wang, Ziwei Zhang, Zehuan Yuan, Hang Li, Wenwu Zhu
Recently, self-supervised learning for graph neural networks (GNNs) has attracted considerable attention because of their notable successes in learning the representation of graph-structure data. However, the formation of a real-world graph typically arises from the highly complex interaction of many latent factors. The existing self-supervised learning methods for GNNs are inherently holistic and neglect the entanglement of the latent factors, resulting in the learned representations suboptimal for downstream tasks and difficult to be interpreted. Learning disentangled graph representations with self-supervised learning poses great challenges and remains largely ignored by the existing literature. In this paper, we introduce the Disentangled Graph Contrastive Learning (DGCL) method, which is able to learn disentangled graph-level representations with self-supervision. In particular, we first identify the latent factors of the input graph and derive its factorized representations. Each of the factorized representations describes a latent and disentangled aspect pertinent to a specific latent factor of the graph. Then we propose a novel factor-wise discrimination objective in a contrastive learning manner, which can force the factorized representations to independently reflect the expressive information from different latent factors. Extensive experiments on both synthetic and real-world datasets demonstrate the superiority of our method against several state-of-the-art baselines.
****

Widening the Pipeline in Human-Guided Reinforcement Learning with Explanation and Context-Aware Data Augmentation
Lin Guan, Mudit Verma, Suna (Sihang) Guo, Ruohan Zhang, Subbarao Kambhampati
Human explanation (e.g., in terms of feature importance) has been recently used to extend the communication channel between human and agent in interactive machine learning. Under this setting, human trainers provide not only the ground truth but also some form of explanation. However, this kind of human guidance was only investigated in supervised learning tasks, and it remains unclear how to best incorporate this type of human knowledge into deep reinforcement learning. In this paper, we present the first study of using human visual explanations in human-in-the-loop reinforcement learning (HIRL). We focus on the task of learning from feedback, in which the human trainer not only gives binary evaluative "good" or "bad" feedback for queried state-action pairs, but also provides a visual explanation by annotating relevant features in images. We propose EXPAND (EXPlanation AugmeNted feeDback) to encourage the model to encode task-relevant features through a context-aware data augmentation that only perturbs irrelevant features in human salient information. We choose five tasks, namely Pixel-Taxi and four Atari games, to evaluate the performance and sample efficiency of this approach. We show that our method significantly outperforms methods leveraging human explanation that are adapted from supervised learning, and Human-in-the-loop RL baselines that only utilize evaluative feedback.
****

SOLQ: Segmenting Objects by Learning Queries
Bin Dong, Fangao Zeng, Tiancai Wang, Xiangyu Zhang, Yichen Wei
In this paper, we propose an end-to-end framework for instance segmentation. Based on the recently introduced DETR, our method, termed SOLQ, segments objects by learning unified queries. In SOLQ, each query represents one object and has multiple representations: class, location and mask. The object queries learned perform classification, box regression and mask encoding simultaneously in an unifie

d vector form. During training phase, the mask vectors encoded are supervised by the compression coding of raw spatial masks. In inference time,mask vectors produced can be directly transformed to spatial masks by the inverse process of compression coding. Experimental results show that SOLQ can achieve state-of-the-art performance, surpassing most of existing approaches. Moreover, the joint learning of unified query representation can greatly improve the detection performance of DETR. We hope our SOLQ can serve as a strong baseline for the Transformer-based instance segmentation.
****
Extending Lagrangian and Hamiltonian Neural Networks with Differentiable Contact Models
Yaofeng Desmond Zhong, Biswadip Dey, Amit Chakraborty
The incorporation of appropriate inductive bias plays a critical role in learning dynamics from data. A growing body of work has been exploring ways to enforce energy conservation in the learned dynamics by encoding Lagrangian or Hamiltonian dynamics into the neural network architecture. These existing approaches are based on differential equations, which do not allow discontinuity in the states and thereby limit the class of systems one can learn. However, in reality, most physical systems, such as legged robots and robotic manipulators, involve contacts and collisions, which introduce discontinuities in the states. In this paper, we introduce a differentiable contact model, which can capture contact mechanics: frictionless/frictional, as well as elastic/inelastic. This model can also accommodate inequality constraints, such as limits on the joint angles. The proposed contact model extends the scope of Lagrangian and Hamiltonian neural networks by allowing simultaneous learning of contact and system properties. We demonstrate this framework on a series of challenging 2D and 3D physical systems with different coefficients of restitution and friction. The learned dynamics can be used as a differentiable physics simulator for downstream gradient-based optimization tasks, such as planning and control.
****
Best-case lower bounds in online learning
Cristóbal Guzmán, Nishant Mehta, Ali Mortazavi
Much of the work in online learning focuses on the study of sublinear upper bounds on the regret. In this work, we initiate the study of best-case lower bounds in online convex optimization, wherein we bound the largest \emph{improvement} an algorithm can obtain relative to the single best action in hindsight. This problem is motivated by the goal of better understanding the adaptivity of a learning algorithm. Another motivation comes from fairness: it is known that best-case lower bounds are instrumental in obtaining algorithms for decision-theoretic online learning (DTOL) that satisfy a notion of group fairness. Our contributions are a general method to provide best-case lower bounds in Follow The Regularized Leader (FTRL) algorithms with time-varying regularizers, which we use to show that best-case lower bounds are of the same order as existing upper regret bounds: this includes situations with a fixed learning rate, decreasing learning rates, timeless methods, and adaptive gradient methods. In stark contrast, we show that the linearized version of FTRL can attain negative linear regret. Finally, in DTOL with two experts and binary losses, we fully characterize the best-case sequences, which provides a finer understanding of the best-case lower bounds.
****
A Comprehensively Tight Analysis of Gradient Descent for PCA
Zhiqiang Xu, 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-authors prior to requesting a name change in the electronic proceedings.
****
On Robust Optimal Transport: Computational Complexity and Barycenter Computation
Khang Le, Huy Nguyen, Quang M Nguyen, Tung Pham, Hung Bui, Nhat Ho
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Asymptotically Best Causal Effect Identification with Multi-Armed Bandits
Alan Malek, Silvia Chiappa
This paper considers the problem of selecting a formula for identifying a causal quantity of interest among a set of available formulas. We assume an online setting in which the investigator may alter the data collection mechanism in a data-dependent way with the aim of identifying the formula with lowest asymptotic variance in as few samples as possible. We formalize this setting by using the best-arm-identification bandit framework where the standard goal of learning the arm with the lowest loss is replaced with the goal of learning the arm that will produce the best estimate. We introduce new tools for constructing finite-sample confidence bounds on estimates of the asymptotic variance that account for the estimation of potentially complex nuisance functions, and adapt the best-arm-identification algorithms of LUCB and Successive Elimination to use these bounds. We validate our method by providing upper bounds on the sample complexity and an empirical study on artificially generated data.
****
Learning rule influences recurrent network representations but not attractor structure in decision-making tasks
Brandon McMahan, Michael Kleinman, Jonathan Kao
Recurrent neural networks (RNNs) are popular tools for studying computational dynamics in neurobiological circuits. However, due to the dizzying array of design choices, it is unclear if computational dynamics unearthed from RNNs provide reliable neurobiological inferences. Understanding the effects of design choices on RNN computation is valuable in two ways. First, invariant properties that persist in RNNs across a wide range of design choices are more likely to be candidate neurobiological mechanisms. Second, understanding what design choices lead to similar dynamical solutions reduces the burden of imposing that all design choices be totally faithful replications of biology. We focus our investigation on how RNN learning rule and task design affect RNN computation. We trained large populations of RNNs with different, but commonly used, learning rules on decision-making tasks inspired by neuroscience literature. For relatively complex tasks, we find that attractor topology is invariant to the choice of learning rule, but representational geometry is not. For simple tasks, we find that attractor topology depends on task input noise. However, when a task becomes increasingly complex, RNN attractor topology becomes invariant to input noise. Together, our results suggest that RNN dynamics are robust across learning rules but can be sensitive to the training task design, especially for simpler tasks.
****
Few-Shot Segmentation via Cycle-Consistent Transformer
Gengwei Zhang, Guoliang Kang, Yi Yang, Yunchao Wei
Few-shot segmentation aims to train a segmentation model that can fast adapt to novel classes with few exemplars. The conventional training paradigm is to learn to make predictions on query images conditioned on the features from support images. Previous methods only utilized the semantic-level prototypes of support images as the conditional information. These methods cannot utilize all pixel-wise support information for the query predictions, which is however critical for the segmentation task. In this paper, we focus on utilizing pixel-wise relationships between support and target images to facilitate the few-shot semantic segmentation task. We design a novel Cycle-Consistent Transformer (CyCTR) module to aggregate pixel-wise support features into query ones. CyCTR performs cross-attention between features from different images, i.e. support and query images. We observe that there may exist unexpected irrelevant pixel-level support features. Directly performing cross-attention may aggregate these features from support to query and bias the query features. Thus, we propose using a novel cycle-consistent attention mechanism to filter out possible harmful support features and encourage query features to attend to the most informative pixels from support images. Experiments on all few-shot segmentation benchmarks demonstrate that our propos

ed CyCTR leads to remarkable improvement compared to previous state-of-the-art methods. Specifically, on Pascal-5^i and COCO-20^i datasets, we achieve 66.6% and 45.6% mIoU for 5-shot segmentation, outperforming previous state-of-the-art by 4.6% and 7.1% respectively.
****
DropGNN: Random Dropouts Increase the Expressiveness of Graph Neural Networks
Pál András Papp, Karolis Martinkus, Lukas Faber, Roger Wattenhofer
This paper studies Dropout Graph Neural Networks (DropGNNs), a new approach that aims to overcome the limitations of standard GNN frameworks. In DropGNNs, we execute multiple runs of a GNN on the input graph, with some of the nodes randomly and independently dropped in each of these runs. Then, we combine the results of these runs to obtain the final result. We prove that DropGNNs can distinguish various graph neighborhoods that cannot be separated by message passing GNNs. We derive theoretical bounds for the number of runs required to ensure a reliable distribution of dropouts, and we prove several properties regarding the expressive capabilities and limits of DropGNNs. We experimentally validate our theoretical findings on expressiveness. Furthermore, we show that DropGNNs perform competitively on established GNN benchmarks.
****
Photonic Differential Privacy with Direct Feedback Alignment
Ruben Ohana, Hamlet Medina, Julien Launay, Alessandro Cappelli, Iacopo Poli, Liva Ralaivola, Alain Rakotomamonjy
Optical Processing Units (OPUs) -- low-power photonic chips dedicated to large scale random projections -- have been used in previous work to train deep neural networks using Direct Feedback Alignment (DFA), an effective alternative to backpropagation. Here, we demonstrate how to leverage the intrinsic noise of optical random projections to build a differentially private DFA mechanism, making OPUs a solution of choice to provide a \emph{private-by-design} training. We provide a theoretical analysis of our adaptive privacy mechanism, carefully measuring how the noise of optical random projections propagates in the process and gives rise to provable Differential Privacy. Finally, we conduct experiments demonstrating the ability of our learning procedure to achieve solid end-task performance.

****
Searching Parameterized AP Loss for Object Detection
Tao Chenxin, Zizhang Li, Xizhou Zhu, Gao Huang, Yong Liu, jifeng dai
Loss functions play an important role in training deep-network-based object detectors. The most widely used evaluation metric for object detection is Average Precision (AP), which captures the performance of localization and classification sub-tasks simultaneously. However, due to the non-differentiable nature of the AP metric, traditional object detectors adopt separate differentiable losses for the two sub-tasks. Such a mis-alignment issue may well lead to performance degradation. To address this, existing works seek to design surrogate losses for the AP metric manually, which requires expertise and may still be sub-optimal. In this paper, we propose Parameterized AP Loss, where parameterized functions are introduced to substitute the non-differentiable components in the AP calculation. Different AP approximations are thus represented by a family of parameterized functions in a unified formula. Automatic parameter search algorithm is then employed to search for the optimal parameters. Extensive experiments on the COCO benchmark with three different object detectors (i.e., RetinaNet, Faster R-CNN, and Deformable DETR) demonstrate that the proposed Parameterized AP Loss consistently outperforms existing handcrafted losses. Code shall be released.
****
Fair Exploration via Axiomatic Bargaining
Jackie Baek, Vivek Farias
Motivated by the consideration of fairly sharing the cost of exploration between multiple groups in learning problems, we develop the Nash bargaining solution in the context of multi-armed bandits. Specifically, the 'grouped' bandit associated with any multi-armed bandit problem associates, with each time step, a single group from some finite set of groups. The utility gained by a given group unde

r some learning policy is naturally viewed as the reduction in that group's regret relative to the regret that group would have incurred 'on its own'. We derive policies that yield the Nash bargaining solution relative to the set of incremental utilities possible under any policy. We show that on the one hand, the 'price of fairness' under such policies is limited, while on the other hand, regret optimal policies are arbitrarily unfair under generic conditions. Our theoretical development is complemented by a case study on contextual bandits for warfarin dosing where we are concerned with the cost of exploration across multiple races and age groups.
****
Unifying lower bounds on prediction dimension of convex surrogates
Jessica Finocchiaro, Rafael Frongillo, Bo Waggoner
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Ultrahyperbolic Neural Networks
Marc Law
Riemannian space forms, such as the Euclidean space, sphere and hyperbolic space, are popular and powerful representation spaces in machine learning. For instance, hyperbolic geometry is appropriate to represent graphs without cycles and has been used to extend Graph Neural Networks. Recently, some pseudo-Riemannian space forms that generalize both hyperbolic and spherical geometries have been exploited to learn a specific type of nonparametric embedding called ultrahyperbolic. The lack of geodesic between every pair of ultrahyperbolic points makes the task of learning parametric models (e.g., neural networks) difficult. This paper introduces a method to learn parametric models in ultrahyperbolic space. We experimentally show the relevance of our approach in the tasks of graph and node classification.
****
NeuroMLR: Robust & Reliable Route Recommendation on Road Networks
Jayant Jain, Vrittika Bagadia, Sahil Manchanda, Sayan Ranu
Predicting the most likely route from a source location to a destination is a core functionality in mapping services. Although the problem has been studied in the literature, two key limitations remain to be addressed. First, our study reveals that a significant portion of the routes recommended by existing methods fail to reach the destination. Second, existing techniques are transductive in nature; hence, they fail to recommend routes if unseen roads are encountered at inference time. In this paper, we address these limitations through an inductive algorithm called NeuroMLR. NeuroMLR learns a generative model from historical trajectories by conditioning on three explanatory factors: the current location, the destination, and real-time traffic conditions. The conditional distributions are learned through a novel combination of Lipschitz embedding with Graph Convolutional Networks (GCN) using historical trajectory data. Through in-depth experiments on real-world datasets, we establish that NeuroMLR imparts significant improvement in accuracy over the state of the art. More importantly, NeuroMLR generalizes dramatically better to unseen data and the recommended routes reach the destination with much higher likelihood than existing techniques.
****
Risk Bounds and Calibration for a Smart Predict-then-Optimize Method
Heyuan Liu, Paul Grigas
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Three-dimensional spike localization and improved motion correction for Neuropixels recordings
Julien Boussard, Erdem Varol, Hyun Dong Lee, Nishchal Dethe, Liam Paninski

Neuropixels (NP) probes are dense linear multi-electrode arrays that have rapidly become essential tools for studying the electrophysiology of large neural populations. Unfortunately, a number of challenges remain in analyzing the large datasets output by these probes. Here we introduce several new methods for extracting useful spiking information from NP probes. First, we use a simple point neuron model, together with a neural-network denoiser, to efficiently map spikes detected on the probe into three-dimensional localizations. Previous methods localized spikes in two dimensions only; we show that the new localization approach is significantly more robust and provides an improved feature set for clustering spikes according to neural identity (``spike sorting"). Next, we apply a Poisson denoising method to the resulting three-dimensional point-cloud representation of the data, and show that the resulting 3D images can be accurately registered over time, leading to improved tracking of time-varying neural activity over the probe, and in turn, crisper estimates of neural clusters over time. The code to reproduce our results and an example neuropixels dataset is provided in the supplementary material.
****
Semi-Supervised Semantic Segmentation via Adaptive Equalization Learning
Hanzhe Hu, Fangyun Wei, Han Hu, Qiwei Ye, Jinshi Cui, Liwei Wang
Due to the limited and even imbalanced data, semi-supervised semantic segmentation tends to have poor performance on some certain categories, e.g., tailed categories in Cityscapes dataset which exhibits a long-tailed label distribution. Existing approaches almost all neglect this problem, and treat categories equally. Some popular approaches such as consistency regularization or pseudo-labeling may even harm the learning of under-performing categories, that the predictions or pseudo labels of these categories could be too inaccurate to guide the learning on the unlabeled data. In this paper, we look into this problem, and propose a novel framework for semi-supervised semantic segmentation, named adaptive equalization learning (AEL). AEL adaptively balances the training of well and badly performed categories, with a confidence bank to dynamically track category-wise performance during training. The confidence bank is leveraged as an indicator to tilt training towards under-performing categories, instantiated in three strategies: 1) adaptive Copy-Paste and CutMix data augmentation approaches which give more chance for under-performing categories to be copied or cut; 2) an adaptive data sampling approach to encourage pixels from under-performing category to be sampled; 3) a simple yet effective re-weighting method to alleviate the training noise raised by pseudo-labeling. Experimentally, AEL outperforms the state-of-the-art methods by a large margin on the Cityscapes and Pascal VOC benchmarks under various data partition protocols. Code is available at https://github.com/hzhupku/SemiSeg-AEL.
****
On the Bias-Variance-Cost Tradeoff of Stochastic Optimization
Yifan Hu, Xin Chen, Niao He
We consider stochastic optimization when one only has access to biased stochastic oracles of the objective, and obtaining stochastic gradients with low biases comes at high costs. This setting captures a variety of optimization paradigms widely used in machine learning, such as conditional stochastic optimization, bilevel optimization, and distributionally robust optimization. We examine a family of multi-level Monte Carlo (MLMC) gradient methods that exploit a delicate trade-off among the bias, the variance, and the oracle cost. We provide a systematic study of their convergences and total computation complexities for strongly convex, convex, and nonconvex objectives, and demonstrate their superiority over the naive biased stochastic gradient method. Moreover, when applied to conditional stochastic optimization, the MLMC gradient methods significantly improve the best-known sample complexity in the literature.
****
Averaging on the Bures-Wasserstein manifold: dimension-free convergence of gradient descent
Jason Altschuler, Sinho Chewi, Patrik R Gerber, Austin Stromme
We study first-order optimization algorithms for computing the barycenter of Gau

ssian distributions with respect to the optimal transport metric. Although the objective is geodesically non-convex, Riemannian gradient descent empirically converges rapidly, in fact faster than off-the-shelf methods such as Euclidean gradient descent and SDP solvers. This stands in stark contrast to the best-known theoretical results, which depend exponentially on the dimension. In this work, we prove new geodesic convexity results which provide stronger control of the iterates, yielding a dimension-free convergence rate. Our techniques also enable the analysis of two related notions of averaging, the entropically-regularized barycenter and the geometric median, providing the first convergence guarantees for these problems.
****

Reinforcement Learning in Newcomblike Environments
James Bell, Linda Linsefors, Caspar Oesterheld, Joar Skalse
Newcomblike decision problems have been studied extensively in the decision theory literature, but they have so far been largely absent in the reinforcement learning literature. In this paper we study value-based reinforcement learning algorithms in the Newcomblike setting, and answer some of the fundamental theoretical questions about the behaviour of such algorithms in these environments. We show that a value-based reinforcement learning agent cannot converge to a policy that is not \emph{ratifiable}, i.e., does not only choose actions that are optimal given that policy. This gives us a powerful tool for reasoning about the limit behaviour of agents -- for example, it lets us show that there are Newcomblike environments in which a reinforcement learning agent cannot converge to any optimal policy. We show that a ratifiable policy always exists in our setting, but that there are cases in which a reinforcement learning agent normally cannot converge to it (and hence cannot converge at all). We also prove several results about the possible limit behaviours of agents in cases where they do not converge to any policy.
****

Comprehensive Knowledge Distillation with Causal Intervention
Xiang Deng, Zhongfei Zhang
Knowledge distillation (KD) addresses model compression by distilling knowledge from a large model (teacher) to a smaller one (student). The existing distillation approaches mainly focus on using different criteria to align the sample representations learned by the student and the teacher, while they fail to transfer the class representations. Good class representations can benefit the sample representation learning by shaping the sample representation distribution. On the other hand, the existing approaches enforce the student to fully imitate the teacher while ignoring the fact that the teacher is typically not perfect. Although the teacher has learned rich and powerful representations, it also contains unignorable bias knowledge which is usually induced by the context prior (e.g., background) in the training data. To address these two issues, in this paper, we propose comprehensive, interventional distillation (CID) that captures both sample and class representations from the teacher while removing the bias with causal intervention. Different from the existing literature that uses the softened logits of the teacher as the training targets, CID considers the softened logits as the context information of an image, which is further used to remove the biased knowledge based on causal inference. Keeping the good representations while removing the bad bias enables CID to have a better generalization ability on test data and a better transferability across different datasets against the existing state-of-the-art approaches, which is demonstrated by extensive experiments on several benchmark datasets.
****

Reinforcement Learning with Latent Flow
Wenling Shang, Xiaofei Wang, Aravind Srinivas, Aravind Rajeswaran, Yang Gao, Pieter Abbeel, Misha Laskin
Temporal information is essential to learning effective policies with Reinforcement Learning (RL). However, current state-of-the-art RL algorithms either assume that such information is given as part of the state space or, when learning from pixels, use the simple heuristic of frame-stacking to implicitly capture temp

oral information present in the image observations. This heuristic is in contras t to the current paradigm in video classification architectures, which utilize e xplicit encodings of temporal information through methods such as optical flow a nd two-stream architectures to achieve state-of-the-art performance. Inspired by leading video classification architectures, we introduce the Flow of Latents fo r Reinforcement Learning (Flare), a network architecture for RL that explicitly encodes temporal information through latent vector differences. We show that Fla re recovers optimal performance in state-based RL without explicit access to the state velocity, solely with positional state information. Flare is the most sam ple efficient model-free pixel-based RL algorithm on the DeepMind Control suite when evaluated on the 500k and 1M step benchmarks across 5 challenging control t asks, and, when used with Rainbow DQN, outperforms the competitive baseline on A tari games at 100M time step benchmark across 8 challenging games.
****

Understanding How Encoder-Decoder Architectures Attend

Kyle Aitken, Vinay Ramasesh, Yuan Cao, Niru Maheswaranathan

Encoder-decoder networks with attention have proven to be a powerful way to solv e many sequence-to-sequence tasks. In these networks, attention aligns encoder a nd decoder states and is often used for visualizing network behavior. However, t he mechanisms used by networks to generate appropriate attention matrices are st ill mysterious. Moreover, how these mechanisms vary depending on the particular architecture used for the encoder and decoder (recurrent, feed-forward, etc.) ar e also not well understood. In this work, we investigate how encoder-decoder net works solve different sequence-to-sequence tasks. We introduce a way of decompos ing hidden states over a sequence into temporal (independent of input) and input -driven (independent of sequence position) components. This reveals how attentio n matrices are formed: depending on the task requirements, networks rely more he avily on either the temporal or input-driven components. These findings hold acr oss both recurrent and feed-forward architectures despite their differences in f orming the temporal components. Overall, our results provide new insight into th e inner workings of attention-based encoder-decoder networks.
****

Latent Execution for Neural Program Synthesis Beyond Domain-Specific Languages

Xinyun Chen, Dawn Song, Yuandong Tian

Program synthesis from input-output (IO) examples has been a long-standing chall enge. While recent works demonstrated limited success on domain-specific languag es (DSL), it remains highly challenging to apply them to real-world programming languages, such as C. Due to complicated syntax and token variation, there are t hree major challenges: (1) unlike many DSLs, programs in languages like C need t o compile first and are not executed via interpreters; (2) the program search sp ace grows exponentially when the syntax and semantics of the programming languag e become more complex; and (3) collecting a large-scale dataset of real-world pr ograms is non-trivial. As a first step to address these challenges, we propose L aSynth and show its efficacy in a restricted-C domain (i.e., C code with tens of tokens, with sequential, branching, loop and simple arithmetic operations but n o library call). More specifically, LaSynth learns the latent representation to approximate the execution of partially generated programs, even if they are inco mplete in syntax (addressing (1)). The learned execution significantly improves the performance of next token prediction over existing approaches, facilitating search (addressing (2)). Finally, once trained with randomly generated ground-tr uth programs and their IO pairs, LaSynth can synthesize more concise programs th at resemble human-written code. Furthermore, retraining our model with these syn thesized programs yields better performance with fewer samples for both Karel an d C program synthesis, indicating the promise of leveraging the learned program synthesizer to improve the dataset quality for input-output program synthesis (a ddressing (3)). When evaluating on whether the program execution outputs match t he IO pairs, LaSynth achieves 55.2% accuracy on generating simple C code with te ns of tokens including loops and branches, outperforming existing approaches wit hout executors by around 20%.
****

Two steps to risk sensitivity
Christopher Gagne, Peter Dayan
Distributional reinforcement learning (RL) – in which agents learn about all the possible long-term consequences of their actions, and not just the expected value – is of great recent interest. One of the most important affordances of a distributional view is facilitating a modern, measured, approach to risk when outcomes are not completely certain. By contrast, psychological and neuroscientific investigations into decision making under risk have utilized a variety of more venerable theoretical models such as prospect theory that lack axiomatically desirable properties such as coherence. Here, we consider a particularly relevant risk measure for modeling human and animal planning, called conditional value-at-risk (CVaR), which quantifies worst-case outcomes (e.g., vehicle accidents or predation). We first adopt a conventional distributional approach to CVaR in a sequential setting and reanalyze the choices of human decision-makers in the well-known two-step task, revealing substantial risk aversion that had been lurking under stickiness and perseveration. We then consider a further critical property of risk sensitivity, namely time consistency, showing alternatives to this form of CVaR that enjoy this desirable characteristic. We use simulations to examine settings in which the various forms differ in ways that have implications for human and animal planning and behavior.
****
DECAF: Generating Fair Synthetic Data Using Causally-Aware Generative Networks
Boris van Breugel, Trent Kyono, Jeroen Berrevoets, Mihaela van der Schaar
Machine learning models have been criticized for reflecting unfair biases in the training data. Instead of solving for this by introducing fair learning algorithms directly, we focus on generating fair synthetic data, such that any downstream learner is fair. Generating fair synthetic data from unfair data - while remaining truthful to the underlying data-generating process (DGP) - is non-trivial. In this paper, we introduce DECAF: a GAN-based fair synthetic data generator for tabular data. With DECAF we embed the DGP explicitly as a structural causal model in the input layers of the generator, allowing each variable to be reconstructed conditioned on its causal parents. This procedure enables inference time debiasing, where biased edges can be strategically removed for satisfying user-defined fairness requirements. The DECAF framework is versatile and compatible with several popular definitions of fairness. In our experiments, we show that DECAF successfully removes undesired bias and - in contrast to existing methods - is capable of generating high-quality synthetic data. Furthermore, we provide theoretical guarantees on the generator's convergence and the fairness of downstream models.
****
EvoGrad: Efficient Gradient-Based Meta-Learning and Hyperparameter Optimization
Ondrej Bohdal, Yongxin Yang, Timothy Hospedales
Gradient-based meta-learning and hyperparameter optimization have seen significant progress recently, enabling practical end-to-end training of neural networks together with many hyperparameters. Nevertheless, existing approaches are relatively expensive as they need to compute second-order derivatives and store a longer computational graph. This cost prevents scaling them to larger network architectures. We present EvoGrad, a new approach to meta-learning that draws upon evolutionary techniques to more efficiently compute hypergradients. EvoGrad estimates hypergradient with respect to hyperparameters without calculating second-order gradients, or storing a longer computational graph, leading to significant improvements in efficiency. We evaluate EvoGrad on three substantial recent meta-learning applications, namely cross-domain few-shot learning with feature-wise transformations, noisy label learning with Meta-Weight-Net and low-resource cross-lingual learning with meta representation transformation. The results show that EvoGrad significantly improves efficiency and enables scaling meta-learning to bigger architectures such as from ResNet10 to ResNet34.
****
Biological learning in key-value memory networks
Danil Tyulmankov, Ching Fang, Annapurna Vadaparty, Guangyu Robert Yang

In neuroscience, classical Hopfield networks are the standard biologically plaus ible model of long-term memory, relying on Hebbian plasticity for storage and at tractor dynamics for recall. In contrast, memory-augmented neural networks in ma chine learning commonly use a key-value mechanism to store and read out memories in a single step. Such augmented networks achieve impressive feats of memory co mpared to traditional variants, yet their biological relevance is unclear. We pr opose an implementation of basic key-value memory that stores inputs using a com bination of biologically plausible three-factor plasticity rules. The same rules are recovered when network parameters are meta-learned. Our network performs on par with classical Hopfield networks on autoassociative memory tasks and can be naturally extended to continual recall, heteroassociative memory, and sequence learning. Our results suggest a compelling alternative to the classical Hopfield network as a model of biological long-term memory.
****
Correlated Stochastic Block Models: Exact Graph Matching with Applications to Re covering Communities
Miklos Racz, Anirudh Sridhar
We consider the task of learning latent community structure from multiple correl ated networks. First, we study the problem of learning the latent vertex corresp ondence between two edge-correlated stochastic block models, focusing on the reg ime where the average degree is logarithmic in the number of vertices. We derive the precise information-theoretic threshold for exact recovery: above the thres hold there exists an estimator that outputs the true correspondence with probabi lity close to 1, while below it no estimator can recover the true correspondence with probability bounded away from 0. As an application of our results, we show how one can exactly recover the latent communities using \emph{multiple} correl ated graphs in parameter regimes where it is information-theoretically impossibl e to do so using just a single graph.
****
Twice regularized MDPs and the equivalence between robustness and regularization
Esther Derman, Matthieu Geist, 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.
****
Nearly Minimax Optimal Reinforcement Learning for Discounted MDPs
Jiafan He, Dongruo Zhou, 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.
****
Sparse Deep Learning: A New Framework Immune to Local Traps and Miscalibration
Yan Sun, Wenjun Xiong, Faming Liang
Deep learning has powered recent successes of artificial intelligence (AI). Howe ver, the deep neural network, as the basic model of deep learning, has suffered from issues such as local traps and miscalibration. In this paper, we provide a new framework for sparse deep learning, which has the above issues addressed in a coherent way. In particular, we lay down a theoretical foundation for sparse deep learning and propose prior annealing algorithms for learning sparse neural networks. The former has successfully tamed the sparse deep neural network into the framework of statistical modeling, enabling prediction uncertainty correctly quantified. The latter can be asymptotically guaranteed to converge to the glob al optimum, enabling the validity of the down-stream statistical inference. Nume rical result indicates the superiority of the proposed method compared to the ex isting ones.
****
Calibrating Predictions to Decisions: A Novel Approach to Multi-Class Calibratio n

Shengjia Zhao, Michael Kim, Roshni Sahoo, Tengyu Ma, Stefano Ermon
When facing uncertainty, decision-makers want predictions they can trust. A machine learning provider can convey confidence to decision-makers by guaranteeing their predictions are distribution calibrated--- amongst the inputs that receive a predicted vector of class probabilities q, the actual distribution over classes is given by q. For multi-class prediction problems, however, directly optimizing predictions under distribution calibration tends to be infeasible, requiring sample complexity that grows exponentially in the number of classes C. In this work, we introduce a new notion---decision calibration---that requires the predicted distribution and true distribution over classes to be ``indistinguishable'' to downstream decision-makers. This perspective gives a new characterization of distribution calibration: a predictor is distribution calibrated if and only if it is decision calibrated with respect to all decision-makers. Our main result shows that under a mild restriction, unlike distribution calibration, decision calibration is actually feasible. We design a recalibration algorithm that provably achieves decision calibration efficiently, provided that the decision-makers have a bounded number of actions (e.g., polynomial in C). We validate our recalibration algorithm empirically: compared to existing methods, decision calibration improves decision-making on skin lesion and ImageNet classification with modern neural network predictors.
****
Lower Bounds and Optimal Algorithms for Smooth and Strongly Convex Decentralized Optimization Over Time-Varying Networks
Dmitry Kovalev, Elnur Gasanov, Alexander Gasnikov, 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.
****
Testing Probabilistic Circuits
Yash Pralhad Pote, Kuldeep S Meel
Probabilistic circuits (PCs) are a powerful modeling framework for representing tractable probability distributions over combinatorial spaces. In machine learning and probabilistic programming, one is often interested in understanding whether the distributions learned using PCs are close to the desired distribution. Thus, given two probabilistic circuits, a fundamental problem of interest is to determine whether their distributions are close to each other.The primary contribution of this paper is a closeness test for PCs with respect to the total variation distance metric. Our algorithm utilizes two common PC queries, counting and sampling. In particular, we provide a poly-time probabilistic algorithm to check the closeness of two PCs, when the PCs support tractable approximate counting and sampling. We demonstrate the practical efficiency of our algorithmic framework via a detailed experimental evaluation of a prototype implementation against a set of 375 PC benchmarks. We find that our test correctly decides the closeness of all 375 PCs within 3600 seconds.
****
Pseudo-Spherical Contrastive Divergence
Lantao Yu, Jiaming Song, Yang Song, Stefano Ermon
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
NORESQA: A Framework for Speech Quality Assessment using Non-Matching References
Pranay Manocha, Buye Xu, Anurag Kumar
The perceptual task of speech quality assessment (SQA) is a challenging task for machines to do. Objective SQA methods that rely on the availability of the corresponding clean reference have been the primary go-to approaches for SQA. Clearly, these methods fail in real-world scenarios where the ground truth clean references are not available. In recent years, non-intrusive methods that train neura

l networks to predict ratings or scores have attracted much attention, but they suffer from several shortcomings such as lack of robustness, reliance on labeled data for training and so on. In this work, we propose a new direction for speech quality assessment. Inspired by human's innate ability to compare and assess the quality of speech signals even when they have non-matching contents, we propose a novel framework that predicts a subjective relative quality score for the given speech signal with respect to any provided reference without using any subjective data. We show that neural networks trained using our framework produce scores that correlate well with subjective mean opinion scores (MOS) and are also competitive to methods such as DNSMOS, which explicitly relies on MOS from humans for training networks. Moreover, our method also provides a natural way to embed quality-related information in neural networks, which we show is helpful for downstream tasks such as speech enhancement.
****
AFEC: Active Forgetting of Negative Transfer in Continual Learning
Liyuan Wang, Mingtian Zhang, Zhongfan Jia, Qian Li, Chenglong Bao, Kaisheng Ma, Jun Zhu, Yi Zhong
Continual learning aims to learn a sequence of tasks from dynamic data distributions. Without accessing to the old training samples, knowledge transfer from the old tasks to each new task is difficult to determine, which might be either positive or negative. If the old knowledge interferes with the learning of a new task, i.e., the forward knowledge transfer is negative, then precisely remembering the old tasks will further aggravate the interference, thus decreasing the performance of continual learning. By contrast, biological neural networks can actively forget the old knowledge that conflicts with the learning of a new experience, through regulating the learning-triggered synaptic expansion and synaptic convergence. Inspired by the biological active forgetting, we propose to actively forget the old knowledge that limits the learning of new tasks to benefit continual learning. Under the framework of Bayesian continual learning, we develop a novel approach named Active Forgetting with synaptic Expansion-Convergence (AFEC). Our method dynamically expands parameters to learn each new task and then selectively combines them, which is formally consistent with the underlying mechanism of biological active forgetting. We extensively evaluate AFEC on a variety of continual learning benchmarks, including CIFAR-10 regression tasks, visual classification tasks and Atari reinforcement tasks, where AFEC effectively improves the learning of new tasks and achieves the state-of-the-art performance in a plug-and-play way.
****
Heterogeneous Multi-player Multi-armed Bandits: Closing the Gap and Generalization
Chengshuai Shi, Wei Xiong, Cong Shen, Jing Yang
Despite the significant interests and many progresses in decentralized multi-player multi-armed bandits (MP-MAB) problems in recent years, the regret gap to the natural centralized lower bound in the heterogeneous MP-MAB setting remains open. In this paper, we propose BEACON -- Batched Exploration with Adaptive COmmunicatioN -- that closes this gap. BEACON accomplishes this goal with novel contributions in implicit communication and efficient exploration. For the former, we propose a novel adaptive differential communication (ADC) design that significantly improves the implicit communication efficiency. For the latter, a carefully crafted batched exploration scheme is developed to enable incorporation of the combinatorial upper confidence bound (CUCB) principle. We then generalize the existing linear-reward MP-MAB problems, where the system reward is always the sum of individually collected rewards, to a new MP-MAB problem where the system reward is a general (nonlinear) function of individual rewards. We extend BEACON to solve this problem and prove a logarithmic regret. BEACON bridges the algorithm design and regret analysis of combinatorial MAB (CMAB) and MP-MAB, two largely disjointed areas in MAB, and the results in this paper suggest that this previously ignored connection is worth further investigation.
****
SWAD: Domain Generalization by Seeking Flat Minima

Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, Sungrae Park
Domain generalization (DG) methods aim to achieve generalizability to an unseen target domain by using only training data from the source domains. Although a variety of DG methods have been proposed, a recent study shows that under a fair evaluation protocol, called DomainBed, the simple empirical risk minimization (ERM) approach works comparable to or even outperforms previous methods. Unfortunately, simply solving ERM on a complex, non-convex loss function can easily lead to sub-optimal generalizability by seeking sharp minima. In this paper, we theoretically show that finding flat minima results in a smaller domain generalization gap. We also propose a simple yet effective method, named Stochastic Weight Averaging Densely (SWAD), to find flat minima. SWAD finds flatter minima and suffers less from overfitting than does the vanilla SWA by a dense and overfit-aware stochastic weight sampling strategy. SWAD shows state-of-the-art performances on five DG benchmarks, namely PACS, VLCS, OfficeHome, TerraIncognita, and DomainNet, with consistent and large margins of +1.6% averagely on out-of-domain accuracy. We also compare SWAD with conventional generalization methods, such as data augmentation and consistency regularization methods, to verify that the remarkable performance improvements are originated from by seeking flat minima, not from better in-domain generalizability. Last but not least, SWAD is readily adaptable to existing DG methods without modification; the combination of SWAD and an existing DG method further improves DG performances. Source code is available at https://github.com/khanrc/swad.
****
Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting
Haixu Wu, Jiehui Xu, Jianmin Wang, Mingsheng Long
Extending the forecasting time is a critical demand for real applications, such as extreme weather early warning and long-term energy consumption planning. This paper studies the long-term forecasting problem of time series. Prior Transformer-based models adopt various self-attention mechanisms to discover the long-range dependencies. However, intricate temporal patterns of the long-term future prohibit the model from finding reliable dependencies. Also, Transformers have to adopt the sparse versions of point-wise self-attentions for long series efficiency, resulting in the information utilization bottleneck. Going beyond Transformers, we design Autoformer as a novel decomposition architecture with an Auto-Correlation mechanism. We break with the pre-processing convention of series decomposition and renovate it as a basic inner block of deep models. This design empowers Autoformer with progressive decomposition capacities for complex time series. Further, inspired by the stochastic process theory, we design the Auto-Correlation mechanism based on the series periodicity, which conducts the dependencies discovery and representation aggregation at the sub-series level. Auto-Correlation outperforms self-attention in both efficiency and accuracy. In long-term forecasting, Autoformer yields state-of-the-art accuracy, with a 38% relative improvement on six benchmarks, covering five practical applications: energy, traffic, economics, weather and disease. Code is available at this repository: https://github.com/thuml/Autoformer.
****
Predicting Event Memorability from Contextual Visual Semantics
Qianli Xu, Fen Fang, Ana Molino, Vigneshwaran Subbaraju, Joo-Hwee Lim
Episodic event memory is a key component of human cognition. Predicting event memorability,i.e., to what extent an event is recalled, is a tough challenge in memory research and has profound implications for artificial intelligence. In this study, we investigate factors that affect event memorability according to a cued recall process. Specifically, we explore whether event memorability is contingent on the event context, as well as the intrinsic visual attributes of image cues. We design a novel experiment protocol and conduct a large-scale experiment with 47 elder subjects over 3 months. Subjects’ memory of life events is tested in a cued recall process. Using advanced visual analytics methods, we build a first-of-its-kind event memorability dataset (called R3) with rich information abo

ut event context and visual semantic features. Furthermore, we propose a contextual event memory network (CEMNet) that tackles multi-modal input to predict item-wise event memorability, which outperforms competitive benchmarks. The findings inform deeper understanding of episodic event memory, and open up a new avenue for prediction of human episodic memory. Source code is available at https://github.com/ffzzy840304/Predicting-Event-Memorability.
****
Achieving Forgetting Prevention and Knowledge Transfer in Continual Learning
Zixuan Ke, Bing Liu, Nianzu Ma, Hu Xu, Lei Shu
Continual learning (CL) learns a sequence of tasks incrementally with the goal of achieving two main objectives: overcoming catastrophic forgetting (CF) and encouraging knowledge transfer (KT) across tasks. However, most existing techniques focus only on overcoming CF and have no mechanism to encourage KT, and thus do not do well in KT. Although several papers have tried to deal with both CF and KT, our experiments show that they suffer from serious CF when the tasks do not have much shared knowledge. Another observation is that most current CL methods do not use pre-trained models, but it has been shown that such models can significantly improve the end task performance. For example, in natural language processing, fine-tuning a BERT-like pre-trained language model is one of the most effective approaches. However, for CL, this approach suffers from serious CF. An interesting question is how to make the best use of pre-trained models for CL. This paper proposes a novel model called CTR to solve these problems. Our experimental results demonstrate the effectiveness of CTR
****
Bandits with many optimal arms
Rianne de Heide, James Cheshire, Pierre Ménard, Alexandra Carpentier
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Combiner: Full Attention Transformer with Sparse Computation Cost
Hongyu Ren, Hanjun Dai, Zihang Dai, Mengjiao Yang, Jure Leskovec, Dale Schuurmans, Bo Dai
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Geometry Processing with Neural Fields
Guandao Yang, Serge Belongie, Bharath Hariharan, Vladlen Koltun
Most existing geometry processing algorithms use meshes as the default shape representation. Manipulating meshes, however, requires one to maintain high quality in the surface discretization. For example, changing the topology of a mesh usually requires additional procedures such as remeshing. This paper instead proposes the use of neural fields for geometry processing. Neural fields can compactly store complicated shapes without spatial discretization. Moreover, neural fields are infinitely differentiable, which allows them to be optimized for objectives that involve higher-order derivatives. This raises the question: can geometry processing be done entirely using neural fields? We introduce loss functions and architectures to show that some of the most challenging geometry processing tasks, such as deformation and filtering, can be done with neural fields. Experimental results show that our methods are on par with the well-established mesh-based methods without committing to a particular surface discretization. Code is available at https://github.com/stevenygd/NFGP.
****
Contextual Recommendations and Low-Regret Cutting-Plane Algorithms
Sreenivas Gollapudi, Guru Guruganesh, Kostas Kollias, Pasin Manurangsi, Renato Leme, Jon Schneider
Requests for name changes in the electronic proceedings will be accepted with no

questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Speech Separation Using an Asynchronous Fully Recurrent Convolutional Neural Network
Xiaolin Hu, Kai Li, Weiyi Zhang, Yi Luo, Jean-Marie Lemercier, Timo Gerkmann
Recent advances in the design of neural network architectures, in particular those specialized in modeling sequences, have provided significant improvements in speech separation performance. In this work, we propose to use a bio-inspired architecture called Fully Recurrent Convolutional Neural Network (FRCNN) to solve the separation task. This model contains bottom-up, top-down and lateral connections to fuse information processed at various time-scales represented by stages. In contrast to the traditional approach updating stages in parallel, we propose to first update the stages one by one in the bottom-up direction, then fuse information from adjacent stages simultaneously and finally fuse information from all stages to the bottom stage together. Experiments showed that this asynchronous updating scheme achieved significantly better results with much fewer parameters than the traditional synchronous updating scheme on speech separation. In addition, the proposed model achieved competitive or better results with high efficiency as compared to other state-of-the-art approaches on two benchmark datasets.
****
Reinforcement Learning Enhanced Explainer for Graph Neural Networks
Caihua Shan, Yifei Shen, Yao Zhang, Xiang Li, Dongsheng Li
Graph neural networks (GNNs) have recently emerged as revolutionary technologies for machine learning tasks on graphs. In GNNs, the graph structure is generally incorporated with node representation via the message passing scheme, making the explanation much more challenging. Given a trained GNN model, a GNN explainer aims to identify a most influential subgraph to interpret the prediction of an instance (e.g., a node or a graph), which is essentially a combinatorial optimization problem over graph. The existing works solve this problem by continuous relaxation or search-based heuristics. But they suffer from key issues such as violation of message passing and hand-crafted heuristics, leading to inferior interpretability. To address these issues, we propose a RL-enhanced GNN explainer, RG-Explainer, which consists of three main components: starting point selection, iterative graph generation and stopping criteria learning. RG-Explainer could construct a connected explanatory subgraph by sequentially adding nodes from the boundary of the current generated graph, which is consistent with the message passing scheme. Further, we design an effective seed locator to select the starting point, and learn stopping criteria to generate superior explanations. Extensive experiments on both synthetic and real datasets show that RG-Explainer outperforms state-of-the-art GNN explainers. Moreover, RG-Explainer can be applied in the inductive setting, demonstrating its better generalization ability.
****
NAS-Bench-x11 and the Power of Learning Curves
Shen Yan, Colin White, Yash Savani, Frank Hutter
While early research in neural architecture search (NAS) required extreme computational resources, the recent releases of tabular and surrogate benchmarks have greatly increased the speed and reproducibility of NAS research. However, two of the most popular benchmarks do not provide the full training information for each architecture. As a result, on these benchmarks it is not possible to evaluate many types of multi-fidelity algorithms, such as learning curve extrapolation, that require evaluating architectures at arbitrary epochs. In this work, we present a method using singular value decomposition and noise modeling to create surrogate benchmarks, NAS-Bench-111, NAS-Bench-311, and NAS-Bench-NLP11, that output the full training information for each architecture, rather than just the final validation accuracy. We demonstrate the power of using the full training information by introducing a learning curve extrapolation framework to modify single-fidelity algorithms, showing that it leads to improvements over popular single-f

idelity algorithms which claimed to be state-of-the-art upon release.
****
Observation-Free Attacks on Stochastic Bandits
Yinglun Xu, Bhuvesh Kumar, Jacob D. Abernethy
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Learning Disentangled Behavior Embeddings
Changhao Shi, Sivan Schwartz, Shahar Levy, Shay Achvat, Maisan Abboud, Amir Ghanayim, Jackie Schiller, Gal Mishne
To understand the relationship between behavior and neural activity, experiments in neuroscience often include an animal performing a repeated behavior such as a motor task. Recent progress in computer vision and deep learning has shown great potential in the automated analysis of behavior by leveraging large and high-quality video datasets. In this paper, we design Disentangled Behavior Embedding (DBE) to learn robust behavioral embeddings from unlabeled, multi-view, high-resolution behavioral videos across different animals and multiple sessions. We further combine DBE with a stochastic temporal model to propose Variational Disentangled Behavior Embedding (VDBE), an end-to-end approach that learns meaningful discrete behavior representations and generates interpretable behavioral videos. Our models learn consistent behavior representations by explicitly disentangling the dynamic behavioral factors (pose) from time-invariant, non-behavioral nuisance factors (context) in a deep autoencoder, and exploit the temporal structures of pose dynamics. Compared to competing approaches, DBE and VDBE enjoy superior performance on downstream tasks such as fine-grained behavioral motif generation and behavior decoding.
****
The Sensory Neuron as a Transformer: Permutation-Invariant Neural Networks for Reinforcement Learning
Yujin Tang, David Ha
In complex systems, we often observe complex global behavior emerge from a collection of agents interacting with each other in their environment, with each individual agent acting only on locally available information, without knowing the full picture. Such systems have inspired development of artificial intelligence algorithms in areas such as swarm optimization and cellular automata. Motivated by the emergence of collective behavior from complex cellular systems, we build systems that feed each sensory input from the environment into distinct, but identical neural networks, each with no fixed relationship with one another. We show that these sensory networks can be trained to integrate information received locally, and through communication via an attention mechanism, can collectively produce a globally coherent policy. Moreover, the system can still perform its task even if the ordering of its inputs is randomly permuted several times during an episode. These permutation invariant systems also display useful robustness and generalization properties that are broadly applicable. Interactive demo and videos of our results: https://attentionneuron.github.io
****
Fast Extra Gradient Methods for Smooth Structured Nonconvex-Nonconcave Minimax Problems
Sucheol Lee, Donghwan 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.
****
Analysis of Sensing Spectral for Signal Recovery under a Generalized Linear Model
Junjie Ma, Ji Xu, Arian Maleki
Requests for name changes in the electronic proceedings will be accepted with 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.
****
Revisiting ResNets: Improved Training and Scaling Strategies
Irwan Bello, William Fedus, Xianzhi Du, Ekin Dogus Cubuk, Aravind Srinivas, Tsun
g-Yi Lin, Jonathon Shlens, Barret Zoph
Novel computer vision architectures monopolize the spotlight, but the impact of
the model architecture is often conflated with simultaneous changes to training
methodology and scaling strategies.Our work revisits the canonical ResNet and st
udies these three aspects in an effort to disentangle them. Perhaps surprisingly
, we find that training and scaling strategies may matter more than architectura
l changes, and further, that the resulting ResNets match recent state-of-the-art
 models. We show that the best performing scaling strategy depends on the traini
ng regime and offer two new scaling strategies: (1) scale model depth in regimes
 where overfitting can occur (width scaling is preferable otherwise); (2) increa
se image resolution more slowly than previously recommended.Using improved train
ing and scaling strategies, we design a family of ResNet architectures, ResNet-R
S, which are 1.7x - 2.7x faster than EfficientNets on TPUs, while achieving simi
lar accuracies on ImageNet. In a large-scale semi-supervised learning setup, Res
Net-RS achieves 86.2% top-1 ImageNet accuracy, while being 4.7x faster than Effi
cientNet-NoisyStudent. The training techniques improve transfer performance on a
 suite of downstream tasks (rivaling state-of-the-art self-supervised algorithms
) and extend to video classification on Kinetics-400. We recommend practitioners
 use these simple revised ResNets as baselines for future research.
****
Sparse Flows: Pruning Continuous-depth Models
Lucas Liebenwein, Ramin Hasani, Alexander Amini, Daniela Rus
Continuous deep learning architectures enable learning of flexible probabilistic
 models for predictive modeling as neural ordinary differential equations (ODEs)
, and for generative modeling as continuous normalizing flows. In this work, we
design a framework to decipher the internal dynamics of these continuous depth m
odels by pruning their network architectures. Our empirical results suggest that
 pruning improves generalization for neural ODEs in generative modeling. We empi
rically show that the improvement is because pruning helps avoid mode-collapse a
nd flatten the loss surface. Moreover, pruning finds efficient neural ODE repres
entations with up to 98% less parameters compared to the original network, witho
ut loss of accuracy. We hope our results will invigorate further research into t
he performance-size trade-offs of modern continuous-depth models.
****
Spectrum-to-Kernel Translation for Accurate Blind Image Super-Resolution
Guangpin Tao, Xiaozhong Ji, Wenzhuo Wang, Shuo Chen, Chuming Lin, Yun Cao, Tong
Lu, Donghao Luo, Ying Tai
Requests for name changes in the electronic proceedings will be accepted with 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 the Rate of Convergence of Regularized Learning in Games: From Bandits and Un
certainty to Optimism and Beyond
Angeliki Giannou, Emmanouil-Vasileios Vlatakis-Gkaragkounis, Panayotis Mertikopo
ulos
In this paper, we examine the convergence rate of a wide range of regularized me
thods for learning in games. To that end, we propose a unified algorithmic templ
ate that we call “follow the generalized leader” (FTGL), and which includes assp
ecial cases the canonical “follow the regularized leader” algorithm, its optimis
tic variants, extra-gradient schemes, and many others. The proposed framework is
 also sufficiently flexible to account for several different feedback models – f
romfull information to bandit feedback. In this general setting, we show that FT
GL algorithms converge locally to strict Nash equilibria at a rate which does no

t depend on the level of uncertainty faced by the players, but only on the geometry of the regularizer near the equilibrium. In particular, we show that algorithms based on entropic regularization – like the exponential weights algorithm – enjoy a linear convergence rate, while Euclidean projection methods converge to equilibrium in a finite number of iterations, even with bandit feedback.
****

SLAPS: Self-Supervision Improves Structure Learning for Graph Neural Networks
Bahare Fatemi, Layla El Asri, Seyed Mehran Kazemi
Graph neural networks (GNNs) work well when the graph structure is provided. However, this structure may not always be available in real-world applications. One solution to this problem is to infer a task-specific latent structure and then apply a GNN to the inferred graph. Unfortunately, the space of possible graph structures grows super-exponentially with the number of nodes and so the task-specific supervision may be insufficient for learning both the structure and the GNN parameters. In this work, we propose the Simultaneous Learning of Adjacency and GNN Parameters with Self-supervision, or SLAPS, a method that provides more supervision for inferring a graph structure through self-supervision. A comprehensive experimental study demonstrates that SLAPS scales to large graphs with hundreds of thousands of nodes and outperforms several models that have been proposed to learn a task-specific graph structure on established benchmarks.
****

Aligning Pretraining for Detection via Object-Level Contrastive Learning
Fangyun Wei, Yue Gao, Zhirong Wu, Han Hu, Stephen Lin
Image-level contrastive representation learning has proven to be highly effective as a generic model for transfer learning. Such generality for transfer learning, however, sacrifices specificity if we are interested in a certain downstream task. We argue that this could be sub-optimal and thus advocate a design principle which encourages alignment between the self-supervised pretext task and the downstream task. In this paper, we follow this principle with a pretraining method specifically designed for the task of object detection. We attain alignment in the following three aspects: 1) object-level representations are introduced via selective search bounding boxes as object proposals; 2) the pretraining network architecture incorporates the same dedicated modules used in the detection pipeline (e.g. FPN); 3) the pretraining is equipped with object detection properties such as object-level translation invariance and scale invariance. Our method, called Selective Object COntrastive learning (SoCo), achieves state-of-the-art results for transfer performance on COCO detection using a Mask R-CNN framework. Code is available at https://github.com/hologerry/SoCo.
****

Double/Debiased Machine Learning for Dynamic Treatment Effects
Greg Lewis, Vasilis Syrgkanis
We consider the estimation of treatment effects in settings when multiple treatments are assigned over time and treatments can have a causal effect on future outcomes. We propose an extension of the double/debiased machine learning framework to estimate the dynamic effects of treatments and apply it to a concrete linear Markovian high-dimensional state space model and to general structural nested mean models. Our method allows the use of arbitrary machine learning methods to control for the high dimensional state, subject to a mean square error guarantee, while still allowing parametric estimation and construction of confidence intervals for the dynamic treatment effect parameters of interest. Our method is based on a sequential regression peeling process, which we show can be equivalently interpreted as a Neyman orthogonal moment estimator. This allows us to show root-n asymptotic normality of the estimated causal effects.
****

Local Disentanglement in Variational Auto-Encoders Using Jacobian L_1 Regularization
Travers Rhodes, Daniel Lee
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-auth

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

Design of Experiments for Stochastic Contextual Linear Bandits
Andrea Zanette, Kefan Dong, Jonathan N Lee, Emma Brunskill
In the stochastic linear contextual bandit setting there exist several minimax procedures for exploration with policies that are reactive to the data being acquired. In practice, there can be a significant engineering overhead to deploy these algorithms, especially when the dataset is collected in a distributed fashion or when a human in the loop is needed to implement a different policy. Exploring with a single non-reactive policy is beneficial in such cases. Assuming some batch contexts are available, we design a single stochastic policy to collect a good dataset from which a near-optimal policy can be extracted. We present a theoretical analysis as well as numerical experiments on both synthetic and real-world datasets.
****

Encoding Spatial Distribution of Convolutional Features for Texture Representation
Yong Xu, Feng Li, Zhile Chen, Jinxiu Liang, Yuhui Quan
Existing convolutional neural networks (CNNs) often use global average pooling (GAP) to aggregate feature maps into a single representation. However, GAP cannot well characterize complex distributive patterns of spatial features while such patterns play an important role in texture-oriented applications, e.g., material recognition and ground terrain classification. In the context of texture representation, this paper addressed the issue by proposing Fractal Encoding (FE), a feature encoding module grounded by multi-fractal geometry. Considering a CNN feature map as a union of level sets of points lying in the 2D space, FE characterizes their spatial layout via a local-global hierarchical fractal analysis which examines the multi-scale power behavior on each level set. This enables a CNN to encode the regularity on the spatial arrangement of image features, leading to a robust yet discriminative spectrum descriptor. In addition, FE has trainable parameters for data adaptivity and can be easily incorporated into existing CNNs for end-to-end training. We applied FE to ResNet-based texture classification and retrieval, and demonstrated its effectiveness on several benchmark datasets.
****

Training Certifiably Robust Neural Networks with Efficient Local Lipschitz Bounds
Yujia Huang, Huan Zhang, Yuanyuan Shi, J. Zico Kolter, Anima Anandkumar
Certified robustness is a desirable property for deep neural networks in safety-critical applications, and popular training algorithms can certify robustness of a neural network by computing a global bound on its Lipschitz constant. However, such a bound is often loose: it tends to over-regularize the neural network and degrade its natural accuracy. A tighter Lipschitz bound may provide a better tradeoff between natural and certified accuracy, but is generally hard to compute exactly due to non-convexity of the network. In this work, we propose an efficient and trainable \emph{local} Lipschitz upper bound by considering the interactions between activation functions (e.g. ReLU) and weight matrices. Specifically, when computing the induced norm of a weight matrix, we eliminate the corresponding rows and columns where the activation function is guaranteed to be a constant in the neighborhood of each given data point, which provides a provably tighter bound than the global Lipschitz constant of the neural network. Our method can be used as a plug-in module to tighten the Lipschitz bound in many certifiable training algorithms. Furthermore, we propose to clip activation functions (e.g., ReLU and MaxMin) with a learnable upper threshold and a sparsity loss to assist the network to achieve an even tighter local Lipschitz bound. Experimentally, we show that our method consistently outperforms state-of-the-art methods in both clean and certified accuracy on MNIST, CIFAR-10 and TinyImageNet datasets with various network architectures.
****

Average-Reward Learning and Planning with Options
Yi Wan, Abhishek Naik, Rich Sutton

We extend the options framework for temporal abstraction in reinforcement learni
ng from discounted Markov decision processes (MDPs) to average-reward MDPs. Our
contributions include general convergent off-policy inter-option learning algori
thms, intra-option algorithms for learning values and models, as well as sample-
based planning variants of our learning algorithms. Our algorithms and convergen
ce proofs extend those recently developed by Wan, Naik, and Sutton. We also exte
nd the notion of option-interrupting behaviour from the discounted to the averag
e-reward formulation. We show the efficacy of the proposed algorithms with exper
iments on a continuing version of the Four-Room domain.
****
SSAL: Synergizing between Self-Training and Adversarial Learning for Domain Adap
tive Object Detection
Muhammad Akhtar Munir, Muhammad Haris Khan, M. Sarfraz, Mohsen Ali
We study adapting trained object detectors to unseen domains manifesting signifi
cant variations of object appearance, viewpoints and backgrounds. Most current m
ethods align domains by either using image or instance-level feature alignment i
n an adversarial fashion. This often suffers due to the presence of unwanted bac
kground and as such lacks class-specific alignment. A common remedy to promote c
lass-level alignment is to use high confidence predictions on the unlabelled dom
ain as pseudo labels. These high confidence predictions are often fallacious sin
ce the model is poorly calibrated under domain shift. In this paper, we propose
to leverage model’s predictive uncertainty to strike the right balance between a
dversarial feature alignment and class-level alignment. Specifically, we measure
predictive uncertainty on class assignments and the bounding box predictions. M
odel predictions with low uncertainty are used to generate pseudo-labels for sel
f-supervision, whereas the ones with higher uncertainty are used to generate til
es for an adversarial feature alignment stage. This synergy between tiling aroun
d the uncertain object regions and generating pseudo-labels from highly certain
object regions allows us to capture both the image and instance level context du
ring the model adaptation stage. We perform extensive experiments covering vario
us domain shift scenarios. Our approach improves upon existing state-of-the-art
methods with visible margins.
****
Counterexample Guided RL Policy Refinement Using Bayesian Optimization
Briti Gangopadhyay, Pallab Dasgupta
Constructing Reinforcement Learning (RL) policies that adhere to safety requirem
ents is an emerging field of study. RL agents learn via trial and error with an
objective to optimize a reward signal. Often policies that are designed to accum
ulate rewards do not satisfy safety specifications. We present a methodology for
counterexample guided refinement of a trained RL policy against a given safety
specification. Our approach has two main components. The first component is an a
pproach to discover failure trajectories using Bayesian optimization over multip
le parameters of uncertainty from a policy learnt in a model-free setting. The s
econd component selectively modifies the failure points of the policy using grad
ient-based updates. The approach has been tested on several RL environments, and
we demonstrate that the policy can be made to respect the safety specifications
through such targeted changes.
****
Stable, Fast and Accurate: Kernelized Attention with Relative Positional Encodin
g
Shengjie Luo, Shanda Li, Tianle Cai, Di He, Dinglan Peng, Shuxin Zheng, Guolin K
e, Liwei Wang, 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.
****
Learning in Non-Cooperative Configurable Markov Decision Processes
Giorgia Ramponi, Alberto Maria Metelli, Alessandro Concetti, Marcello Restelli
The Configurable Markov Decision Process framework includes two entities: a Rein

forcement Learning agent and a configurator that can modify some environmental parameters to improve the agent's performance. This presupposes that the two actors have the same reward functions. What if the configurator does not have the same intentions as the agent? This paper introduces the Non-Cooperative Configurable Markov Decision Process, a setting that allows having two (possibly different) reward functions for the configurator and the agent. Then, we consider an online learning problem, where the configurator has to find the best among a finite set of possible configurations. We propose two learning algorithms to minimize the configurator's expected regret, which exploits the problem's structure, depending on the agent's feedback. While a naive application of the UCB algorithm yields a regret that grows indefinitely over time, we show that our approach suffers only bounded regret. Furthermore, we empirically show the performance of our algorithm in simulated domains.
****

Identification of Partially Observed Linear Causal Models: Graphical Conditions for the Non-Gaussian and Heterogeneous Cases
Jeffrey Adams, Niels Hansen, Kun Zhang
In causal discovery, linear non-Gaussian acyclic models (LiNGAMs) have been studied extensively. While the causally sufficient case is well understood, in many real problems the observed variables are not causally related. Rather, they are generated by latent variables, such as confounders and mediators, which may themselves be causally related. Existing results on the identification of the causal structure among the latent variables often require very strong graphical assumptions. In this paper, we consider partially observed linear models with either non-Gaussian or heterogeneous errors. In that case we give two graphical conditions which are necessary for identification of the causal structure. These conditions are closely related to sparsity of the causal edges. Together with one additional condition on the coefficients, which holds generically for any graph, the two graphical conditions are also sufficient for identifiability. These new conditions can be satisfied even when there is a large number of latent variables. We demonstrate the validity of our results on synthetic data.
****

DIB-R++: Learning to Predict Lighting and Material with a Hybrid Differentiable Renderer
Wenzheng Chen, Joey Litalien, Jun Gao, Zian Wang, Clement Fuji Tsang, Sameh Khamis, Or Litany, Sanja Fidler
We consider the challenging problem of predicting intrinsic object properties from a single image by exploiting differentiable renderers. Many previous learning-based approaches for inverse graphics adopt rasterization-based renderers and assume naive lighting and material models, which often fail to account for non-Lambertian, specular reflections commonly observed in the wild. In this work, we propose DIBR++, a hybrid differentiable renderer which supports these photorealistic effects by combining rasterization and ray-tracing, taking the advantage of their respective strengths---speed and realism. Our renderer incorporates environmental lighting and spatially-varying material models to efficiently approximate light transport, either through direct estimation or via spherical basis functions. Compared to more advanced physics-based differentiable renderers leveraging path tracing, DIBR++ is highly performant due to its compact and expressive shading model, which enables easy integration with learning frameworks for geometry, reflectance and lighting prediction from a single image without requiring any ground-truth. We experimentally demonstrate that our approach achieves superior material and lighting disentanglement on synthetic and real data compared to existing rasterization-based approaches and showcase several artistic applications including material editing and relighting.
****

Coresets for Time Series Clustering
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.
****
A Variational Perspective on Diffusion-Based Generative Models and Score Matching
Chin-Wei Huang, Jae Hyun Lim, Aaron C. Courville
Discrete-time diffusion-based generative models and score matching methods have shown promising results in modeling high-dimensional image data. Recently, Song et al. (2021) show that diffusion processes that transform data into noise can be reversed via learning the score function, i.e. the gradient of the log-density of the perturbed data. They propose to plug the learned score function into an inverse formula to define a generative diffusion process. Despite the empirical success, a theoretical underpinning of this procedure is still lacking. In this work, we approach the (continuous-time) generative diffusion directly and derive a variational framework for likelihood estimation, which includes continuous-time normalizing flows as a special case, and can be seen as an infinitely deep variational autoencoder. Under this framework, we show that minimizing the score-matching loss is equivalent to maximizing a lower bound of the likelihood of the plug-in reverse SDE proposed by Song et al. (2021), bridging the theoretical gap.
****
Online Active Learning with Surrogate Loss Functions
Giulia DeSalvo, Claudio Gentile, Tobias Sommer Thune
We derive a novel active learning algorithm in the streaming setting for binary classification tasks. The algorithm leverages weak labels to minimize the number of label requests, and trains a model to optimize a surrogate loss on a resulting set of labeled and weak-labeled points. Our algorithm jointly admits two crucial properties: theoretical guarantees in the general agnostic setting and a strong empirical performance. Our theoretical analysis shows that the algorithm attains favorable generalization and label complexity bounds, while our empirical study on 18 real-world datasets demonstrate that the algorithm outperforms standard baselines, including the Margin Algorithm, or Uncertainty Sampling, a high-performing active learning algorithm favored by practitioners.
****
Does Preprocessing Help Training Over-parameterized Neural Networks?
Zhao Song, Shuo Yang, Ruizhe 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.
****
Causal Influence Detection for Improving Efficiency in Reinforcement Learning
Maximilian Seitzer, Bernhard Schölkopf, Georg Martius
Many reinforcement learning (RL) environments consist of independent entities that interact sparsely. In such environments, RL agents have only limited influence over other entities in any particular situation. Our idea in this work is that learning can be efficiently guided by knowing when and what the agent can influence with its actions. To achieve this, we introduce a measure of situation-dependent causal influence based on conditional mutual information and show that it can reliably detect states of influence. We then propose several ways to integrate this measure into RL algorithms to improve exploration and off-policy learning. All modified algorithms show strong increases in data efficiency on robotic manipulation tasks.
****
LADA: Look-Ahead Data Acquisition via Augmentation for Deep Active Learning
Yoon-Yeong Kim, Kyungwoo Song, JoonHo Jang, Il-chul Moon
Active learning effectively collects data instances for training deep learning models when the labeled dataset is limited and the annotation cost is high. Data augmentation is another effective technique to enlarge the limited amount of labeled instances. The scarcity of labeled dataset leads us to consider the integration of data augmentation and active learning. One possible approach is a pipeli

ned combination, which selects informative instances via the acquisition function and generates virtual instances from the selected instances via augmentation. However, this pipelined approach would not guarantee the informativeness of the virtual instances. This paper proposes Look-Ahead Data Acquisition via augmentation, or LADA framework, that looks ahead the effect of data augmentation in the process of acquisition. LADA jointly considers both 1) unlabeled data instance to be selected and 2) virtual data instance to be generated by data augmentation, to construct the acquisition function. Moreover, to generate maximally informative virtual instances, LADA optimizes the data augmentation policy to maximize the predictive acquisition score, resulting in the proposal of InfoSTN and InfoMixup. The experimental results of LADA show a significant improvement over the recent augmentation and acquisition baselines that were independently applied.
****
Policy Optimization in Adversarial MDPs: Improved Exploration via Dilated Bonuses
Haipeng Luo, Chen-Yu Wei, Chung-Wei Lee
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Multiclass versus Binary Differentially Private PAC Learning
Satchit Sivakumar, Mark Bun, Marco Gaboardi
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 Robust Change Point Detection
Mengchu Li, Yi 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.
****
Cycle Self-Training for Domain Adaptation
Hong Liu, Jianmin Wang, Mingsheng Long
Mainstream approaches for unsupervised domain adaptation (UDA) learn domain-invariant representations to narrow the domain shift, which are empirically effective but theoretically challenged by the hardness or impossibility theorems. Recently, self-training has been gaining momentum in UDA, which exploits unlabeled target data by training with target pseudo-labels. However, as corroborated in this work, under distributional shift, the pseudo-labels can be unreliable in terms of their large discrepancy from target ground truth. In this paper, we propose Cycle Self-Training (CST), a principled self-training algorithm that explicitly enforces pseudo-labels to generalize across domains. CST cycles between a forward step and a reverse step until convergence. In the forward step, CST generates target pseudo-labels with a source-trained classifier. In the reverse step, CST trains a target classifier using target pseudo-labels, and then updates the shared representations to make the target classifier perform well on the source data. We introduce the Tsallis entropy as a confidence-friendly regularization to improve the quality of target pseudo-labels. We analyze CST theoretically under realistic assumptions, and provide hard cases where CST recovers target ground truth, while both invariant feature learning and vanilla self-training fail. Empirical results indicate that CST significantly improves over the state-of-the-arts on visual recognition and sentiment analysis benchmarks.
****
Novel Visual Category Discovery with Dual Ranking Statistics and Mutual Knowledge Distillation
Bingchen Zhao, Kai Han

In this paper, we tackle the problem of novel visual category discovery, i.e., g rouping unlabelled images from new classes into different semantic partitions by leveraging a labelled dataset that contains images from other different but rel evant categories. This is a more realistic and challenging setting than convent ional semi-supervised learning. We propose a two-branch learning framework for t his problem, with one branch focusing on local part-level information and the ot her branch focusing on overall characteristics. To transfer knowledge from the l abelled data to the unlabelled, we propose using dual ranking statistics on both branches to generate pseudo labels for training on the unlabelled data. We furt her introduce a mutual knowledge distillation method to allow information exchan ge and encourage agreement between the two branches for discovering new categori es, allowing our model to enjoy the benefits of global and local features. We co mprehensively evaluate our method on public benchmarks for generic object classi fication, as well as the more challenging datasets for fine-grained visual recog nition, achieving state-of-the-art performance.
****
Stochastic Anderson Mixing for Nonconvex Stochastic Optimization
Fuchao Wei, Chenglong Bao, Yang Liu
Anderson mixing (AM) is an acceleration method for fixed-point iterations. Despi te its success and wide usage in scientific computing, the convergence theory of AM remains unclear, and its applications to machine learning problems are not w ell explored. In this paper, by introducing damped projection and adaptive regul arization to the classical AM, we propose a Stochastic Anderson Mixing (SAM) sch eme to solve nonconvex stochastic optimization problems. Under mild assumptions, we establish the convergence theory of SAM, including the almost sure convergen ce to stationary points and the worst-case iteration complexity. Moreover, the c omplexity bound can be improved when randomly choosing an iterate as the output. To further accelerate the convergence, we incorporate a variance reduction tech nique into the proposed SAM. We also propose a preconditioned mixing strategy fo r SAM which can empirically achieve faster convergence or better generalization ability. Finally, we apply the SAM method to train various neural networks inclu ding the vanilla CNN, ResNets, WideResNet, ResNeXt, DenseNet and LSTM. Experimen tal results on image classification and language model demonstrate the advantage s of our method.
****
Sample-Efficient Reinforcement Learning for Linearly-Parameterized MDPs with a G enerative Model
Bingyan Wang, Yuling Yan, Jianqing Fan
Requests for name changes in the electronic proceedings will be accepted with 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.
****
NN-Baker: A Neural-network Infused Algorithmic Framework for Optimization Proble ms on Geometric Intersection Graphs
Evan McCarty, Qi Zhao, Anastasios Sidiropoulos, Yusu Wang
Recent years have witnessed a surge of approaches to use neural networks to help tackle combinatorial optimization problems, including graph optimization proble ms. However, theoretical understanding of such approaches remains limited. In th is paper, we consider the geometric setting, where graphs are induced by points in a fixed dimensional Euclidean space. We show that several graph optimization problems can be approximated by an algorithm that is polynomial in graph size n via a framework we propose, call the Baker-paradigm. More importantly, a key adv antage of the Baker-paradigm is that it decomposes the input problem into (at mo st linear number of) small sub-problems of fixed sizes (independent of the size of the input). For the family of such fixed-size sub-problems, we can now design neural networks with universal approximation guarantees to solve them. This lea ds to a mixed algorithmic-ML framework, which we call NN-Baker that has the capa city to approximately solve a family of graph optimization problems (e.g, maximu m independent set and minimum vertex cover) in time linear to input graph size,

and only polynomial to approximation parameter. We instantiate our NN-Baker by a CNN version and GNN version, and demonstrate the effectiveness and efficiency of our approach via a range of experiments.
****

A Note on Sparse Generalized Eigenvalue Problem
Yunfeng Cai, Guanhua Fang, Ping Li
The sparse generalized eigenvalue problem (SGEP) aims to find the leading eigenvector with sparsity structure. SGEP plays an important role in statistical learning and has wide applications including, but not limited to, sparse principal component analysis, sparse canonical correlation analysis and sparse Fisher discriminant analysis, etc. Due to the sparsity constraint, the solution of SGEP entails interesting properties from both numerical and statistical perspectives. In this paper, we provide a detailed sensitivity analysis for SGEP and establish the rate-optimal perturbation bound under the sparse setting. Specifically, we show that the bound is related to the perturbation/noise level and the recovery of the true support of the leading eigenvector as well. We also investigate the estimator of SGEP via imposing a non-convex regularization. Such estimator can achieve the optimal error rate and can recover the sparsity structure as well. Extensive numerical experiments corroborate our theoretical findings via using alternating direction method of multipliers (ADMM)-based computational method.
****

RMIX: Learning Risk-Sensitive Policies for Cooperative Reinforcement Learning Agents
Wei Qiu, Xinrun Wang, Runsheng Yu, Rundong Wang, Xu He, Bo An, Svetlana Obraztsova, Zinovi Rabinovich
Current value-based multi-agent reinforcement learning methods optimize individual Q values to guide individuals' behaviours via centralized training with decentralized execution (CTDE). However, such expected, i.e., risk-neutral, Q value is not sufficient even with CTDE due to the randomness of rewards and the uncertainty in environments, which causes the failure of these methods to train coordinating agents in complex environments. To address these issues, we propose RMIX, a novel cooperative MARL method with the Conditional Value at Risk (CVaR) measure over the learned distributions of individuals' Q values. Specifically, we first learn the return distributions of individuals to analytically calculate CVaR for decentralized execution. Then, to handle the temporal nature of the stochastic outcomes during executions, we propose a dynamic risk level predictor for risk level tuning. Finally, we optimize the CVaR policies with CVaR values used to estimate the target in TD error during centralized training and the CVaR values are used as auxiliary local rewards to update the local distribution via Quantile Regression loss. Empirically, we show that our method outperforms many state-of-the-art methods on various multi-agent risk-sensitive navigation scenarios and challenging StarCraft II cooperative tasks, demonstrating enhanced coordination and revealing improved sample efficiency.
****

Optimal Policies Tend To Seek Power
Alex Turner, Logan Smith, Rohin Shah, Andrew Critch, Prasad Tadepalli
Some researchers speculate that intelligent reinforcement learning (RL) agents would be incentivized to seek resources and power in pursuit of the objectives we specify for them. Other researchers point out that RL agents need not have human-like power-seeking instincts. To clarify this discussion, we develop the first formal theory of the statistical tendencies of optimal policies. In the context of Markov decision processes, we prove that certain environmental symmetries are sufficient for optimal policies to tend to seek power over the environment. These symmetries exist in many environments in which the agent can be shut down or destroyed. We prove that in these environments, most reward functions make it optimal to seek power by keeping a range of options available and, when maximizing average reward, by navigating towards larger sets of potential terminal states.
****

Catalytic Role Of Noise And Necessity Of Inductive Biases In The Emergence Of Co

mpositional Communication
■ukasz Kuci■ski, Tomasz Korbak, Pawe■ Ko■odziej, Piotr Mi■o■
Communication is compositional if complex signals can be represented as a combination of simpler subparts. In this paper, we theoretically show that inductive biases on both the training framework and the data are needed to develop a compositional communication. Moreover, we prove that compositionality spontaneously arises in the signaling games, where agents communicate over a noisy channel. We experimentally confirm that a range of noise levels, which depends on the model and the data, indeed promotes compositionality. Finally, we provide a comprehensive study of this dependence and report results in terms of recently studied compositionality metrics: topographical similarity, conflict count, and context independence.
****
PLUR: A Unifying, Graph-Based View of Program Learning, Understanding, and Repair
Zimin Chen, Vincent J Hellendoorn, Pascal Lamblin, Petros Maniatis, Pierre-Antoine Manzagol, Daniel Tarlow, Subhodeep Moitra
Machine learning for understanding and editing source code has recently attracted significant interest, with many developments in new models, new code representations, and new tasks.This proliferation can appear disparate and disconnected, making each approach seemingly unique and incompatible, thus obscuring the core machine learning challenges and contributions.In this work, we demonstrate that the landscape can be significantly simplified by taking a general approach of mapping a graph to a sequence of tokens and pointers.Our main result is to show that 16 recently published tasks of different shapes can be cast in this form, based on which a single model architecture achieves near or above state-of-the-art results on nearly all tasks, outperforming custom models like code2seq and alternative generic models like Transformers.This unification further enables multi-task learning and a series of cross-cutting experiments about the importance of different modeling choices for code understanding and repair tasks.The full framework, called PLUR, is easily extensible to more tasks, and will be open-sourced (https://github.com/google-research/plur).
****
COCO-LM: Correcting and Contrasting Text Sequences for Language Model Pretraining
Yu Meng, Chenyan Xiong, Payal Bajaj, saurabh tiwary, Paul Bennett, Jiawei Han, XIA SONG
We present a self-supervised learning framework, COCO-LM, that pretrains Language Models by COrrecting and COntrasting corrupted text sequences. Following ELECTRA-style pretraining, COCO-LM employs an auxiliary language model to corrupt text sequences, upon which it constructs two new tasks for pretraining the main model. The first token-level task, Corrective Language Modeling, is to detect and correct tokens replaced by the auxiliary model, in order to better capture token-level semantics. The second sequence-level task, Sequence Contrastive Learning, is to align text sequences originated from the same source input while ensuring uniformity in the representation space. Experiments on GLUE and SQuAD demonstrate that COCO-LM not only outperforms recent state-of-the-art pretrained models in accuracy, but also improves pretraining efficiency. It achieves the MNLI accuracy of ELECTRA with 50% of its pretraining GPU hours. With the same pretraining steps of standard base/large-sized models, COCO-LM outperforms the previous best models by 1+ GLUE average points.
****
Minibatch and Momentum Model-based Methods for Stochastic Weakly Convex Optimization
Qi Deng, Wenzhi Gao
Stochastic model-based methods have received increasing attention lately due to their appealing robustness to the stepsize selection and provable efficiency guarantee. We make two important extensions for improving model-based methods on stochastic weakly convex optimization. First, we propose new minibatch model- based methods by involving a set of samples to approximate the model function in eac

h iteration. For the first time, we show that stochastic algorithms achieve linear speedup over the batch size even for non-smooth and non-convex (particularly, weakly convex) problems. To this end, we develop a novel sensitivity analysis of the proximal mapping involved in each algorithm iteration. Our analysis appears to be of independent interests in more general settings. Second, motivated by the success of momentum stochastic gradient descent, we propose a new stochastic extrapolated model-based method, greatly extending the classic Polyak momentum technique to a wider class of stochastic algorithms for weakly convex optimization. The rate of convergence to some natural stationarity condition is established over a fairly flexible range of extrapolation terms.While mainly focusing on weakly convex optimization, we also extend our work to convex optimization. We apply the minibatch and extrapolated model-based methods to stochastic convex optimization, for which we provide a new complexity bound and promising linear speedup in batch size. Moreover, an accelerated model-based method based on Nesterov's momentum is presented, for which we establish an optimal complexity bound for reaching optimality.
****
XDO: A Double Oracle Algorithm for Extensive-Form Games
Stephen McAleer, JB Lanier, Kevin A Wang, Pierre Baldi, Roy Fox
Policy Space Response Oracles (PSRO) is a reinforcement learning (RL) algorithm for two-player zero-sum games that has been empirically shown to find approximate Nash equilibria in large games. Although PSRO is guaranteed to converge to an approximate Nash equilibrium and can handle continuous actions, it may take an exponential number of iterations as the number of information states (infostates) grows. We propose Extensive-Form Double Oracle (XDO), an extensive-form double oracle algorithm for two-player zero-sum games that is guaranteed to converge to an approximate Nash equilibrium linearly in the number of infostates. Unlike PSRO, which mixes best responses at the root of the game, XDO mixes best responses at every infostate. We also introduce Neural XDO (NXDO), where the best response is learned through deep RL. In tabular experiments on Leduc poker, we find that XDO achieves an approximate Nash equilibrium in a number of iterations an order of magnitude smaller than PSRO. Experiments on a modified Leduc poker game and Oshi-Zumo show that tabular XDO achieves a lower exploitability than CFR with the same amount of computation. We also find that NXDO outperforms PSRO and NFSP on a sequential multidimensional continuous-action game. NXDO is the first deep RL method that can find an approximate Nash equilibrium in high-dimensional continuous-action sequential games.
****
Active Assessment of Prediction Services as Accuracy Surface Over Attribute Combinations
Vihari Piratla, Soumen Chakrabarti, Sunita Sarawagi
Our goal is to evaluate the accuracy of a black-box classification model, not as a single aggregate on a given test data distribution, but as a surface over a large number of combinations of attributes characterizing multiple test data distributions. Such attributed accuracy measures become important as machine learning models get deployed as a service, where the training data distribution is hidden from clients, and different clients may be interested in diverse regions of the data distribution. We present Attributed Accuracy Assay (AAA) --- a Gaussian Process (GP)-based probabilistic estimator for such an accuracy surface. Each attribute combination, called an 'arm' is associated with a Beta density from which the service's accuracy is sampled. We expect the GP to smooth the parameters of the Beta density over related arms to mitigate sparsity. We show that obvious application of GPs cannot address the challenge of heteroscedastic uncertainty over a huge attribute space that is sparsely and unevenly populated. In response, we present two enhancements: pooling sparse observations, and regularizing the scale parameter of the Beta densities. After introducing these innovations, we establish the effectiveness of AAA both in terms of its estimation accuracy and exploration efficiency, through extensive experiments and analysis.
****
A mechanistic multi-area recurrent network model of decision-making

Michael Kleinman, Chandramouli Chandrasekaran, Jonathan Kao
Recurrent neural networks (RNNs) trained on neuroscience-based tasks have been widely used as models for cortical areas performing analogous tasks. However, very few tasks involve a single cortical area, and instead require the coordination of multiple brain areas. Despite the importance of multi-area computation, there is a limited understanding of the principles underlying such computation. We propose to use multi-area RNNs with neuroscience-inspired architecture constraints to derive key features of multi-area computation. In particular, we show that incorporating multiple areas and Dale's Law is critical for biasing the networks to learn biologically plausible solutions. Additionally, we leverage the full observability of the RNNs to show that output-relevant information is preferentially propagated between areas. These results suggest that cortex uses modular computation to generate minimal sufficient representations of task information. More broadly, our results suggest that constrained multi-area RNNs can produce experimentally testable hypotheses for computations that occur within and across multiple brain areas, enabling new insights into distributed computation in neural systems.
****
Learning to Compose Visual Relations
Nan Liu, Shuang Li, Yilun Du, Josh Tenenbaum, Antonio Torralba
The visual world around us can be described as a structured set of objects and their associated relations. An image of a room may be conjured given only the description of the underlying objects and their associated relations. While there has been significant work on designing deep neural networks which may compose individual objects together, less work has been done on composing the individual relations between objects. A principal difficulty is that while the placement of objects is mutually independent, their relations are entangled and dependent on each other. To circumvent this issue, existing works primarily compose relations by utilizing a holistic encoder, in the form of text or graphs. In this work, we instead propose to represent each relation as an unnormalized density (an energy-based model), enabling us to compose separate relations in a factorized manner. We show that such a factorized decomposition allows the model to both generate and edit scenes that have multiple sets of relations more faithfully. We further show that decomposition enables our model to effectively understand the underlying relational scene structure.
****
Identity testing for Mallows model
Róbert Busa-Fekete, Dimitris Fotakis, Balazs Szorenyi, Emmanouil Zampetakis
In this paper, we devise identity tests for ranking data that is generated from Mallows model both in the \emph{asymptotic} and \emph{non-asymptotic} settings. First we consider the case when the central ranking is known, and devise two algorithms for testing the spread parameter of the Mallows model. The first one is obtained by constructing a Uniformly Most Powerful Unbiased (UMPU) test in the asymptotic setting and then converting it into a sample-optimal non-asymptotic identity test. The resulting test is, however, impractical even for medium sized data, because it requires computing the distribution of the sufficient statistic. The second non-asymptotic test is derived from an optimal learning algorithm for the Mallows model. This test is both easy to compute and is sample-optimal for a wide range of parameters. Next, we consider testing Mallows models for the unknown central ranking case. This case can be tackled in the asymptotic setting by introducing a bias that exponentially decays with the sample size. We support all our findings with extensive numerical experiments and show that the proposed tests scale gracefully with the number of items to be ranked.
****
Bandits with Knapsacks beyond the Worst Case
Karthik Abinav Sankararaman, Aleksandrs Slivkins
Bandits with Knapsacks (BwK) is a general model for multi-armed bandits under supply/budget constraints. While worst-case regret bounds for BwK are well-understood, we present three results that go beyond the worst-case perspective. First, we provide upper and lower bounds which amount to a full characterization for lo

garithmic, instance-dependent regret rates.Second, we consider "simple regret" in BwK, which tracks algorithm's performance in a given round, and prove that it is small in all but a few rounds. Third, we provide a "generalreduction" from BwK to bandits which takes advantage of some known helpful structure, and apply this reduction to combinatorial semi-bandits, linear contextual bandits, and multinomial-logit bandits. Our results build on the BwK algorithm from prior work, providing new analyses thereof.
****

Closing the loop in medical decision support by understanding clinical decision-making: A case study on organ transplantation
Yuchao Qin, Fergus Imrie, Alihan Hüyük, Daniel Jarrett, alexander gimson, Mihaela van der Schaar
Significant effort has been placed on developing decision support tools to improve patient care. However, drivers of real-world clinical decisions in complex medical scenarios are not yet well-understood, resulting in substantial gaps between these tools and practical applications. In light of this, we highlight that more attention on understanding clinical decision-making is required both to elucidate current clinical practices and to enable effective human-machine interactions. This is imperative in high-stakes scenarios with scarce available resources. Using organ transplantation as a case study, we formalize the desiderata of methods for understanding clinical decision-making. We show that most existing machine learning methods are insufficient to meet these requirements and propose iTransplant, a novel data-driven framework to learn the factors affecting decisions on organ offers in an instance-wise fashion directly from clinical data, as a possible solution. Through experiments on real-world liver transplantation data from OPTN, we demonstrate the use of iTransplant to: (1) discover which criteria are most important to clinicians for organ offer acceptance; (2) identify patient-specific organ preferences of clinicians allowing automatic patient stratification; and (3) explore variations in transplantation practices between different transplant centers. Finally, we emphasize that the insights gained by iTransplant can be used to inform the development of future decision support tools.
****

Change Point Detection via Multivariate Singular Spectrum Analysis
Arwa Alanqary, Abdullah Alomar, 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.
****

Meta-learning to Improve Pre-training
Aniruddh Raghu, Jonathan Lorraine, Simon Kornblith, Matthew McDermott, David K. Duvenaud
Pre-training (PT) followed by fine-tuning (FT) is an effective method for training neural networks, and has led to significant performance improvements in many domains. PT can incorporate various design choices such as task and data reweighting strategies, augmentation policies, and noise models, all of which can significantly impact the quality of representations learned. The hyperparameters introduced by these strategies therefore must be tuned appropriately. However, setting the values of these hyperparameters is challenging. Most existing methods either struggle to scale to high dimensions, are too slow and memory-intensive, or cannot be directly applied to the two-stage PT and FT learning process. In this work, we propose an efficient, gradient-based algorithm to meta-learn PT hyperparameters. We formalize the PT hyperparameter optimization problem and propose a novel method to obtain PT hyperparameter gradients by combining implicit differentiation and backpropagation through unrolled optimization. We demonstrate that our method improves predictive performance on two real-world domains. First, we optimize high-dimensional task weighting hyperparameters for multitask pre-training on protein-protein interaction graphs and improve AUROC by up to 3.9%. Second, we optimize a data augmentation neural network for self-supervised PT with SimCLR on electrocardiography data and improve AUROC by up to 1.9%.

****
Fair Sparse Regression with Clustering: An Invex Relaxation for a Combinatorial Problem
Adarsh Barik, Jean Honorio
In this paper, we study the problem of fair sparse regression on a biased datase
t where bias depends upon a hidden binary attribute. The presence of a hidden at
tribute adds an extra layer of complexity to the problem by combining sparse reg
ression and clustering with unknown binary labels. The corresponding optimizatio
n problem is combinatorial, but we propose a novel relaxation of it as an invex
optimization problem. To the best of our knowledge, this is the first invex rela
xation for a combinatorial problem. We show that the inclusion of the debiasing/
fairness constraint in our model has no adverse effect on the performance. Rathe
r, it enables the recovery of the hidden attribute. The support of our recovered
regression parameter vector matches exactly with the true parameter vector. Mor
eover, we simultaneously solve the clustering problem by recovering the exact va
lue of the hidden attribute for each sample. Our method uses carefully construct
ed primal dual witnesses to provide theoretical guarantees for the combinatorial
problem. To that end, we show that the sample complexity of our method is logar
ithmic in terms of the dimension of the regression parameter vector.
****
Probabilistic Margins for Instance Reweighting in Adversarial Training
qizhou wang, Feng Liu, Bo Han, Tongliang Liu, Chen Gong, Gang Niu, Mingyuan Zhou
, Masashi Sugiyama
Reweighting adversarial data during training has been recently shown to improve
adversarial robustness, where data closer to the current decision boundaries are
regarded as more critical and given larger weights. However, existing methods m
easuring the closeness are not very reliable: they are discrete and can take onl
y a few values, and they are path-dependent, i.e., they may change given the sam
e start and end points with different attack paths. In this paper, we propose th
ree types of probabilistic margin (PM), which are continuous and path-independen
t, for measuring the aforementioned closeness and reweighing adversarial data. S
pecifically, a PM is defined as the difference between two estimated class-poste
rior probabilities, e.g., such a probability of the true label minus the probabi
lity of the most confusing label given some natural data. Though different PMs c
apture different geometric properties, all three PMs share a negative correlatio
n with the vulnerability of data: data with larger/smaller PMs are safer/riskier
and should have smaller/larger weights. Experiments demonstrated that PMs are r
eliable and PM-based reweighting methods outperformed state-of-the-art counterpa
rts.
****
Unbalanced Optimal Transport through Non-negative Penalized Linear Regression
Laetitia Chapel, Rémi Flamary, Haoran Wu, Cédric Févotte, Gilles Gasso
This paper addresses the problem of Unbalanced Optimal Transport (UOT) in which
the marginal conditions are relaxed (using weighted penalties in lieu of equalit
y) and no additional regularization is enforced on the OT plan. In this context,
we show that the corresponding optimization problem can be reformulated as a no
n-negative penalized linear regression problem. This reformulation allows us to
propose novel algorithms inspired from inverse problems and nonnegative matrix f
actorization. In particular, we consider majorization-minimization which leads i
n our setting to efficient multiplicative updates for a variety of penalties. Fu
rthermore, we derive for the first time an efficient algorithm to compute the re
gularization path of UOT with quadratic penalties. The proposed algorithm provid
es a continuity of piece-wise linear OT plans converging to the solution of bala
nced OT (corresponding to infinite penalty weights). We perform several numerica
l experiments on simulated and real data illustrating the new algorithms, and pr
ovide a detailed discussion about more sophisticated optimization tools that can
further be used to solve OT problems thanks to our reformulation.
****
The Difficulty of Passive Learning in Deep Reinforcement Learning
Georg Ostrovski, Pablo Samuel Castro, Will Dabney

Learning to act from observational data without active environmental interaction is a well-known challenge in Reinforcement Learning (RL). Recent approaches involve constraints on the learned policy or conservative updates, preventing strong deviations from the state-action distribution of the dataset. Although these methods are evaluated using non-linear function approximation, theoretical justifications are mostly limited to the tabular or linear cases. Given the impressive results of deep reinforcement learning, we argue for a need to more clearly understand the challenges in this setting.In the vein of Held & Hein's classic 1963 experiment, we propose the "tandem learning" experimental paradigm which facilitates our empirical analysis of the difficulties in offline reinforcement learning. We identify function approximation in conjunction with fixed data distributions as the strongest factors, thereby extending but also challenging hypotheses stated in past work. Our results provide relevant insights for offline deep reinforcement learning, while also shedding new light on phenomena observed in the online case of learning control.
****
Intriguing Properties of Vision Transformers
Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang
Vision transformers (ViT) have demonstrated impressive performance across numerous machine vision tasks. These models are based on multi-head self-attention mechanisms that can flexibly attend to a sequence of image patches to encode contextual cues. An important question is how such flexibility (in attending image-wide context conditioned on a given patch) can facilitate handling nuisances in natural images e.g., severe occlusions, domain shifts, spatial permutations, adversarial and natural perturbations. We systematically study this question via an extensive set of experiments encompassing three ViT families and provide comparisons with a high-performing convolutional neural network (CNN). We show and analyze the following intriguing properties of ViT: (a)Transformers are highly robust to severe occlusions, perturbations and domain shifts, e.g., retain as high as 60% top-1 accuracy on ImageNet even after randomly occluding 80% of the image content. (b)The robustness towards occlusions is not due to texture bias, instead we show that ViTs are significantly less biased towards local textures, compared to CNNs. When properly trained to encode shape-based features, ViTs demonstrate shape recognition capability comparable to that of human visual system, previously unmatched in the literature. (c)Using ViTs to encode shape representation leads to an interesting consequence of accurate semantic segmentation without pixel-level supervision. (d)Off-the-shelf features from a single ViT model can be combined to create a feature ensemble, leading to high accuracy rates across a range of classification datasets in both traditional and few-shot learning paradigms. We show effective features of ViTs are due to flexible and dynamic receptive fields possible via self-attention mechanisms. Our code will be publicly released.
****
PartialFed: Cross-Domain Personalized Federated Learning via Partial Initialization
Benyuan Sun, Hongxing Huo, YI YANG, Bo Bai
The burst of applications empowered by massive data have aroused unprecedented privacy concerns in AI society. Currently, data confidentiality protection has been one core issue during deep model training. Federated Learning (FL), which enables privacy-preserving training across multiple silos, gained rising popularity for its parameter-only communication. However, previous works have shown that FL revealed a significant performance drop if the data distributions are heterogeneous among different clients, especially when the clients have cross-domain characteristic, such as traffic, aerial and in-door. To address this challenging problem, we propose a novel idea, PartialFed, which loads a subset of the global model’s parameters rather than loading the entire model used in most previous works. We first validate our algorithm with manually decided loading strategies inspired by various expert priors, named PartialFed-Fix. Then we develop PartialFed-Adaptive, which automatically selects personalized loading strategy for each cl

ient. The superiority of our algorithm is proved by demonstrating the new state-
of-the-art results on cross-domain federated classification and detection. In pa
rticular, solely by initializing a small fraction of layers locally, we improve
the performance of FedAvg on Office-Home and UODB by 4.88% and 2.65%, respective
ly. Further studies show that the adaptive strategy performs significantly bette
r on domains with large deviation, e.g. improves AP50 by 4.03% and 4.89% on aeri
al and medical image detection compared to FedAvg.
****
Adaptive Diffusion in Graph Neural Networks
Jialin Zhao, Yuxiao Dong, Ming Ding, Evgeny Kharlamov, Jie Tang
The success of graph neural networks (GNNs) largely relies on the process of agg
regating information from neighbors defined by the input graph structures. Notab
ly, message passing based GNNs, e.g., graph convolutional networks, leverage the
 immediate neighbors of each node during the aggregation process, and recently,
graph diffusion convolution (GDC) is proposed to expand the propagation neighbor
hood by leveraging generalized graph diffusion. However, the neighborhood size i
n GDC is manually tuned for each graph by conducting grid search over the valida
tion set, making its generalization practically limited. To address this issue,
we propose the adaptive diffusion convolution (ADC) strategy to automatically le
arn the optimal neighborhood size from the data. Furthermore, we break the conve
ntional assumption that all GNN layers and feature channels (dimensions) should
use the same neighborhood for propagation. We design strategies to enable ADC to
 learn a dedicated propagation neighborhood for each GNN layer and each feature
channel, making the GNN architecture fully coupled with graph structures---the u
nique property that differs GNNs from traditional neural networks. By directly p
lugging ADC into existing GNNs, we observe consistent and significant outperform
ance over both GDC and their vanilla versions across various datasets, demonstra
ting the improved model capacity brought by automatically learning unique neighb
orhood size per layer and per channel in GNNs.
****
Recurrent Submodular Welfare and Matroid Blocking Semi-Bandits
Orestis Papadigenopoulos, 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.
****
Representer Point Selection via Local Jacobian Expansion for Post-hoc Classifier
 Explanation of Deep Neural Networks and Ensemble Models
Yi Sui, Ga Wu, Scott Sanner
Requests for name changes in the electronic proceedings will be accepted with 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.
****
Editing a classifier by rewriting its prediction rules
Shibani Santurkar, Dimitris Tsipras, Mahalaxmi Elango, David Bau, Antonio Torral
ba, Aleksander Madry
We propose a methodology for modifying the behavior of a classifier by directly
rewriting its prediction rules. Our method requires virtually no additional data
 collection and can be applied to a variety of settings, including adapting a mo
del to new environments, and modifying it to ignore spurious features.
****
How Modular should Neural Module Networks Be for Systematic Generalization?
Vanessa D'Amario, Tomotake Sasaki, Xavier Boix
Neural Module Networks (NMNs) aim at Visual Question Answering (VQA) via composi
tion of modules that tackle a sub-task. NMNs are a promising strategy to achieve
 systematic generalization, i.e., overcoming biasing factors in the training dis
tribution. However, the aspects of NMNs that facilitate systematic generalizatio
n are not fully understood. In this paper, we demonstrate that the degree of mod

ularity of the NMN have large influence on systematic generalization. In a serie
s of experiments on three VQA datasets (VQA-MNIST, SQOOP, and CLEVR-CoGenT), our
 results reveal that tuning the degree of modularity, especially at the image en
coder stage, reaches substantially higher systematic generalization. These findi
ngs lead to new NMN architectures that outperform previous ones in terms of syst
ematic generalization.
****
Contrast and Mix: Temporal Contrastive Video Domain Adaptation with Background M
ixing
Aadarsh Sahoo, Rutav Shah, Rameswar Panda, Kate Saenko, Abir Das
Unsupervised domain adaptation which aims to adapt models trained on a labeled s
ource domain to a completely unlabeled target domain has attracted much attentio
n in recent years. While many domain adaptation techniques have been proposed fo
r images, the problem of unsupervised domain adaptation in videos remains largel
y underexplored. In this paper, we introduce Contrast and Mix (CoMix), a new con
trastive learning framework that aims to learn discriminative invariant feature
representations for unsupervised video domain adaptation. First, unlike existing
 methods that rely on adversarial learning for feature alignment, we utilize tem
poral contrastive learning to bridge the domain gap by maximizing the similarity
 between encoded representations of an unlabeled video at two different speeds a
s well as minimizing the similarity between different videos played at different
 speeds. Second, we propose a novel extension to the temporal contrastive loss b
y using background mixing that allows additional positives per anchor, thus adap
ting contrastive learning to leverage action semantics shared across both domain
s. Moreover, we also integrate a supervised contrastive learning objective using
 target pseudo-labels to enhance discriminability of the latent space for video
domain adaptation. Extensive experiments on several benchmark datasets demonstra
te the superiority of our proposed approach over state-of-the-art methods. Proje
ct page: https://cvir.github.io/projects/comix.
****
The Flip Side of the Reweighted Coin: Duality of Adaptive Dropout and Regulariza
tion
Daniel LeJeune, Hamid Javadi, Richard Baraniuk
Among the most successful methods for sparsifying deep (neural) networks are tho
se that adaptively mask the network weights throughout training. By examining th
is masking, or dropout, in the linear case, we uncover a duality between such ad
aptive methods and regularization through the so-called “η-trick” that casts bot
h as iteratively reweighted optimizations. We show that any dropout strategy tha
t adapts to the weights in a monotonic way corresponds to an effective subquadra
tic regularization penalty, and therefore leads to sparse solutions. We obtain t
he effective penalties for several popular sparsification strategies, which are
remarkably similar to classical penalties commonly used in sparse optimization.
Considering variational dropout as a case study, we demonstrate similar empirica
l behavior between the adaptive dropout method and classical methods on the task
 of deep network sparsification, validating our theory.
****
Active Learning of Convex Halfspaces on Graphs
Maximilian Thiessen, Thomas Gaertner
We systematically study the query complexity of learning geodesically convex hal
fspaces on graphs. Geodesic convexity is a natural generalisation of Euclidean c
onvexity and allows the definition of convex sets and halfspaces on graphs. We p
rove an upper bound on the query complexity linear in the treewidth and the mini
mum hull set size but only logarithmic in the diameter. We show tight lower boun
ds along well-established separation axioms and identify the Radon number as a c
entral parameter of the query complexity and the VC dimension. While previous bo
unds typically depend on the cut size of the labelling, all parameters in our bo
unds can be computed from the unlabelled graph. We provide evidence that ground-
truth communities in real-world graphs are often convex and empirically compare
our proposed approach with other active learning algorithms.
****

Differentiable Spike: Rethinking Gradient-Descent for Training Spiking Neural Ne
tworks
Yuhang Li, Yufei Guo, Shanghang Zhang, Shikuang Deng, Yongqing Hai, Shi Gu
Spiking Neural Networks (SNNs) have emerged as a biology-inspired method mimicki
ng the spiking nature of brain neurons. This bio-mimicry derives SNNs' energy ef
ficiency of inference on neuromorphic hardware. However, it also causes an intri
nsic disadvantage in training high-performing SNNs from scratch since the discre
te spike prohibits the gradient calculation. To overcome this issue, the surroga
te gradient (SG) approach has been proposed as a continuous relaxation. Yet the
heuristic choice of SG leaves it vacant how the SG benefits the SNN training. In
 this work, we first theoretically study the gradient descent problem in SNN tra
ining and introduce finite difference gradient to quantitatively analyze the tra
ining behavior of SNN. Based on the introduced finite difference gradient, we pr
opose a new family of Differentiable Spike (Dspike) functions that can adaptivel
y evolve during training to find the optimal shape and smoothness for gradient e
stimation. Extensive experiments over several popular network structures show th
at training SNN with Dspike consistently outperforms the state-of-the-art traini
ng methods. For example, on the CIFAR10-DVS classification task, we can train a
spiking ResNet-18 and achieve 75.4% top-1 accuracy with 10 time steps.
****
Probabilistic Entity Representation Model for Reasoning over Knowledge Graphs
Nurendra Choudhary, Nikhil Rao, Sumeet Katariya, Karthik Subbian, Chandan Reddy
Logical reasoning over Knowledge Graphs (KGs) is a fundamental technique that ca
n provide an efficient querying mechanism over large and incomplete databases. C
urrent approaches employ spatial geometries such as boxes to learn query represe
ntations that encompass the answer entities and model the logical operations of
projection and intersection. However, their geometry is restrictive and leads to
 non-smooth strict boundaries, which further results in ambiguous answer entitie
s. Furthermore, previous works propose transformation tricks to handle unions wh
ich results in non-closure and, thus, cannot be chained in a stream. In this pap
er, we propose a Probabilistic Entity Representation Model (PERM) to encode enti
ties as a Multivariate Gaussian density with mean and covariance parameters to c
apture its semantic position and smooth decision boundary, respectively. Additio
nally, we also define the closed logical operations of projection, intersection,
 and union that can be aggregated using an end-to-end objective function. On the
 logical query reasoning problem, we demonstrate that the proposed PERM signific
antly outperforms the state-of-the-art methods on various public benchmark KG da
tasets on standard evaluation metrics. We also evaluate PERM’s competence on a C
OVID-19 drug-repurposing case study and show that our proposed work is able to r
ecommend drugs with substantially better F1 than current methods. Finally, we de
monstrate the working of our PERM’s query answering process through a low-dimens
ional visualization of the Gaussian representations.
****
Black Box Probabilistic Numerics
Onur Teymur, Christopher Foley, Philip Breen, Toni Karvonen, Chris J. Oates
Probabilistic numerics casts numerical tasks, such the numerical solution of dif
ferential equations, as inference problems to be solved. One approach is to mode
l the unknown quantity of interest as a random variable, and to constrain this v
ariable using data generated during the course of a traditional numerical method
. However, data may be nonlinearly related to the quantity of interest, renderin
g the proper conditioning of random variables difficult and limiting the range o
f numerical tasks that can be addressed. Instead, this paper proposes to constru
ct probabilistic numerical methods based only on the final output from a traditi
onal method. A convergent sequence of approximations to the quantity of interest
 constitute a dataset, from which the limiting quantity of interest can be extra
polated, in a probabilistic analogue of Richardson’s deferred approach to the li
mit. This black box approach (1) massively expands the range of tasks to which p
robabilistic numerics can be applied, (2) inherits the features and performance
of state-of-the-art numerical methods, and (3) enables provably higher orders of
 convergence to be achieved. Applications are presented for nonlinear ordinary a

nd partial differential equations, as well as for eigenvalue problems—a setting for which no probabilistic numerical methods have yet been developed.
****

Interpolation can hurt robust generalization even when there is no noise
Konstantin Donhauser, Alexandru Tifrea, Michael Aerni, Reinhard Heckel, Fanny Yang
Numerous recent works show that overparameterization implicitly reduces variance for min-norm interpolators and max-margin classifiers. These findings suggest that ridge regularization has vanishing benefits in high dimensions. We challenge this narrative by showing that, even in the absence of noise, avoiding interpolation through ridge regularization can significantly improve generalization. We prove this phenomenon for the robust risk of both linear regression and classification, and hence provide the first theoretical result on \emph{robust overfitting}.
****

On the Equivalence between Neural Network and Support Vector Machine
Yilan Chen, Wei Huang, Lam Nguyen, Tsui-Wei Weng
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 Semantic Representations to Verify Hardware Designs
Shobha Vasudevan, Wenjie (Joe) Jiang, David Bieber, Rishabh Singh, hamid shojaei, C. Richard Ho, Charles Sutton
Verification is a serious bottleneck in the industrial hardware design cycle, routinely requiring person-years of effort. Practical verification relies on a "best effort" process that simulates the design on test inputs. This suggests a new research question: Can this simulation data be exploited to learn a continuous representation of a hardware design that allows us to predict its functionality? As a first approach to this new problem, we introduce Design2Vec, a deep architecture that learns semantic abstractions of hardware designs. The key idea is to work at a higher level of abstraction than the gate or the bit level, namely the Register Transfer Level (RTL), which is somewhat analogous to software source code, and can be represented by a graph that incorporates control and data flow. This allows us to learn representations of RTL syntax and semantics using a graph neural network. We apply these representations to several tasks within verification, including predicting what cover points of the design will be exercised by a test, and generating new tests that will exercise desired cover points. We evaluate Design2Vec on three real-world hardware designs, including an industrial chip used in commercial data centers. Our results demonstrate that Design2Vec dramatically outperforms baseline approaches that do not incorporate the RTL semantics, scales to industrial designs, and can generate tests that exercise design points that are currently hard to cover with manually written tests by design verification experts.
****

Rebooting ACGAN: Auxiliary Classifier GANs with Stable Training
Minguk Kang, Woohyeon Shim, Minsu Cho, Jaesik Park
Conditional Generative Adversarial Networks (cGAN) generate realistic images by incorporating class information into GAN. While one of the most popular cGANs is an auxiliary classifier GAN with softmax cross-entropy loss (ACGAN), it is widely known that training ACGAN is challenging as the number of classes in the dataset increases. ACGAN also tends to generate easily classifiable samples with a lack of diversity. In this paper, we introduce two cures for ACGAN. First, we identify that gradient exploding in the classifier can cause an undesirable collapse in early training, and projecting input vectors onto a unit hypersphere can resolve the problem. Second, we propose the Data-to-Data Cross-Entropy loss (D2D-CE) to exploit relational information in the class-labeled dataset. On this foundation, we propose the Rebooted Auxiliary Classifier Generative Adversarial Network (ReACGAN). The experimental results show that ReACGAN achieves state-of-the-a

rt generation results on CIFAR10, Tiny-ImageNet, CUB200, and ImageNet datasets. We also verify that ReACGAN benefits from differentiable augmentations and that D2D-CE harmonizes with StyleGAN2 architecture. Model weights and a software pack age that provides implementations of representative cGANs and all experiments in our paper are available at https://github.com/POSTECH-CVLab/PyTorch-StudioGAN.
****
Towards a Theoretical Framework of Out-of-Distribution Generalization
Haotian Ye, Chuanlong Xie, Tianle Cai, Ruichen Li, Zhenguo Li, Liwei Wang
Generalization to out-of-distribution (OOD) data is one of the central problems in modern machine learning. Recently, there is a surge of attempts to propose al gorithms that mainly build upon the idea of extracting invariant features. Altho ugh intuitively reasonable, theoretical understanding of what kind of invariance can guarantee OOD generalization is still limited, and generalization to arbitr ary out-of-distribution is clearly impossible. In this work, we take the first s tep towards rigorous and quantitative definitions of 1) what is OOD; and 2) what does it mean by saying an OOD problem is learnable. We also introduce a new con cept of expansion function, which characterizes to what extent the variance is a mplified in the test domains over the training domains, and therefore give a qua ntitative meaning of invariant features. Based on these, we prove an OOD general ization error bound. It turns out that OOD generalization largely depends on the expansion function. As recently pointed out by Gulrajani & Lopez-Paz (2020), an y OOD learning algorithm without a model selection module is incomplete. Our the ory naturally induces a model selection criterion. Extensive experiments on benc hmark OOD datasets demonstrate that our model selection criterion has a signific ant advantage over baselines.
****
Slice Sampling Reparameterization Gradients
David Zoltowski, Diana Cai, Ryan P. Adams
Many probabilistic modeling problems in machine learning use gradient-based opti mization in which the objective takes the form of an expectation. These problems can be challenging when the parameters to be optimized determine the probabilit y distribution under which the expectation is being taken, as the na\"ive Monte Carlo procedure is not differentiable. Reparameterization gradients make it poss ible to efficiently perform optimization of these Monte Carlo objectives by tran sforming the expectation to be differentiable, but the approach is typically lim ited to distributions with simple forms and tractable normalization constants. H ere we describe how to differentiate samples from slice sampling to compute \tex tit{slice sampling reparameterization gradients}, enabling a richer class of Mon te Carlo objective functions to be optimized. Slice sampling is a Markov chain M onte Carlo algorithm for simulating samples from probability distributions; it o nly requires a density function that can be evaluated point-wise up to a normali zation constant, making it applicable to a variety of inference problems and unn ormalized models. Our approach is based on the observation that when the slice e ndpoints are known, the sampling path is a deterministic and differentiable func tion of the pseudo-random variables, since the algorithm is rejection-free. We e valuate the method on synthetic examples and apply it to a variety of applicatio ns with reparameterization of unnormalized probability distributions.
****
Multi-Label Learning with Pairwise Relevance Ordering
Ming-Kun Xie, Sheng-Jun Huang
Precisely annotating objects with multiple labels is costly and has become a cri tical bottleneck in real-world multi-label classification tasks. Instead, decidi ng the relative order of label pairs is obviously less laborious than collecting exact labels. However, the supervised information of pairwise relevance orderin g is less informative than exact labels. It is thus an important challenge to ef fectively learn with such weak supervision. In this paper, we formalize this pro blem as a novel learning framework, called multi-label learning with pairwise re levance ordering (PRO). We show that the unbiased estimator of classification ri sk can be derived with a cost-sensitive loss only from PRO examples. Theoretical ly, we provide the estimation error bound for the proposed estimator and further

prove that it is consistent with respective to the commonly used ranking loss. Empirical studies on multiple datasets and metrics validate the effectiveness of the proposed method.
****
Sampling with Trusthworthy Constraints: A Variational Gradient Framework
Xingchao Liu, Xin Tong, Qiang Liu
Sampling-based inference and learning techniques, especially Bayesian inference, provide an essential approach to handling uncertainty in machine learning (ML). As these techniques are increasingly used in daily life, it becomes essential to safeguard the ML systems with various trustworthy-related constraints, such as fairness, safety, interpretability. Mathematically, enforcing these constraints in probabilistic inference can be cast into sampling from intractable distributions subject to general nonlinear constraints, for which practical efficient algorithms are still largely missing. In this work, we propose a family of constrained sampling algorithms which generalize Langevin Dynamics (LD) and Stein Variational Gradient Descent (SVGD) to incorporate a moment constraint specified by a general nonlinear function. By exploiting the gradient flow structure of LD and SVGD, we derive two types of algorithms for handling constraints, including a primal-dual gradient approach and the constraint controlled gradient descent approach. We investigate the continuous-time mean-field limit of these algorithms and show that they have O(1/t) convergence under mild conditions. Moreover, the LD variant converges linearly assuming that a log Sobolev like inequality holds. Various numerical experiments are conducted to demonstrate the efficiency of our algorithms in trustworthy settings.
****
Robust and Decomposable Average Precision for Image Retrieval
Elias Ramzi, Nicolas THOME, Clément Rambour, Nicolas Audebert, Xavier Bitot
In image retrieval, standard evaluation metrics rely on score ranking, e.g. average precision (AP). In this paper, we introduce a method for robust and decomposable average precision (ROADMAP) addressing two major challenges for end-to-end training of deep neural networks with AP: non-differentiability and non-decomposability.Firstly, we propose a new differentiable approximation of the rank function, which provides an upper bound of the AP loss and ensures robust training. Secondly, we design a simple yet effective loss function to reduce the decomposability gap between the AP in the whole training set and its averaged batch approximation, for which we provide theoretical guarantees.Extensive experiments conducted on three image retrieval datasets show that ROADMAP outperforms several recent AP approximation methods and highlight the importance of our two contributions. Finally, using ROADMAP for training deep models yields very good performances, outperforming state-of-the-art results on the three datasets.Code and instructions to reproduce our results will be made publicly available at https://github.com/elias-ramzi/ROADMAP.
****
Fast rates for prediction with limited expert advice
El Mehdi Saad, Gilles Blanchard
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Probabilistic Transformer For Time Series Analysis
Binh Tang, David S Matteson
Generative modeling of multivariate time series has remained challenging partly due to the complex, non-deterministic dynamics across long-distance timesteps. In this paper, we propose deep probabilistic methods that combine state-space models (SSMs) with transformer architectures. In contrast to previously proposed SSMs, our approaches use attention mechanism to model non-Markovian dynamics in the latent space and avoid recurrent neural networks entirely. We also extend our models to include several layers of stochastic variables organized in a hierarchy for further expressiveness. Compared to transformer models, ours are probabili

stic, non-autoregressive, and capable of generating diverse long-term forecasts with uncertainty estimates. Extensive experiments show that our models consistently outperform competitive baselines on various tasks and datasets, including time series forecasting and human motion prediction.
****
A Hierarchical Reinforcement Learning Based Optimization Framework for Large-scale Dynamic Pickup and Delivery Problems
Yi Ma, Xiaotian Hao, Jianye Hao, Jiawen Lu, Xing Liu, Tong Xialiang, Mingxuan Yuan, Zhigang Li, Jie Tang, Zhaopeng Meng
The Dynamic Pickup and Delivery Problem (DPDP) is an essential problem in the logistics domain, which is NP-hard. The objective is to dynamically schedule vehicles among multiple sites to serve the online generated orders such that the overall transportation cost could be minimized. The critical challenge of DPDP is the orders are not known a priori, i.e., the orders are dynamically generated in real-time. To address this problem, existing methods partition the overall DPDP into fixed-size sub-problems by caching online generated orders and solve each sub-problem, or on this basis to utilize the predicted future orders to optimize each sub-problem further. However, the solution quality and efficiency of these methods are unsatisfactory, especially when the problem scale is very large. In this paper, we propose a novel hierarchical optimization framework to better solve large-scale DPDPs. Specifically, we design an upper-level agent to dynamically partition the DPDP into a series of sub-problems with different scales to optimize vehicles routes towards globally better solutions. Besides, a lower-level agent is designed to efficiently solve each sub-problem by incorporating the strengths of classical operational research-based methods with reinforcement learning-based policies. To verify the effectiveness of the proposed framework, real historical data is collected from the order dispatching system of Huawei Supply Chain Business Unit and used to build a functional simulator. Extensive offline simulation and online testing conducted on the industrial order dispatching system justify the superior performance of our framework over existing baselines.
****
Spatio-Temporal Variational Gaussian Processes
Oliver Hamelijnck, William Wilkinson, Niki Loppi, Arno Solin, Theodoros Damoulas
We introduce a scalable approach to Gaussian process inference that combines spatio-temporal filtering with natural gradient variational inference, resulting in a non-conjugate GP method for multivariate data that scales linearly with respect to time. Our natural gradient approach enables application of parallel filtering and smoothing, further reducing the temporal span complexity to be logarithmic in the number of time steps. We derive a sparse approximation that constructs a state-space model over a reduced set of spatial inducing points, and show that for separable Markov kernels the full and sparse cases exactly recover the standard variational GP, whilst exhibiting favourable computational properties. To further improve the spatial scaling we propose a mean-field assumption of independence between spatial locations which, when coupled with sparsity and parallelisation, leads to an efficient and accurate method for large spatio-temporal problems.
****
MERLOT: Multimodal Neural Script Knowledge Models
Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, Yejin Choi
As humans, we understand events in the visual world contextually, performing multimodal reasoning across time to make inferences about the past, present, and future. We introduce MERLOT, a model that learns multimodal script knowledge by watching millions of YouTube videos with transcribed speech -- in an entirely label-free, self-supervised manner. By pretraining with a mix of both frame-level (spatial) and video-level (temporal) objectives, our model not only learns to match images to temporally corresponding words, but also to contextualize what is happening globally over time. As a result, MERLOT exhibits strong out-of-the-box representations of temporal commonsense, and achieves state-of-the-art performance on 12 different video QA datasets when finetuned. It also transfers well to th

e world of static images, allowing models to reason about the dynamic context behind visual scenes. On Visual Commonsense Reasoning, MERLOT~answers questions correctly with 80.6\% accuracy, outperforming state-of-the-art models of similar size by over 3\%, even those that make heavy use of auxiliary supervised data (like object bounding boxes).Ablation analyses demonstrate the complementary importance of: 1) training on videos versus static images; 2) scaling the magnitude and diversity of the pretraining video corpus; and 3) using diverse objectives that encourage full-stack multimodal reasoning, from the recognition to cognition level.
****
Fast Approximate Dynamic Programming for Infinite-Horizon Markov Decision Processes
Mohamad Amin Sharifi Kolarijani, Gyula Max, Peyman Mohajerin Mohajerin Esfahani
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 Risk Minimization: Learning to Adapt to Domain Shift
Marvin Zhang, Henrik Marklund, Nikita Dhawan, Abhishek Gupta, Sergey Levine, Chelsea Finn
A fundamental assumption of most machine learning algorithms is that the training and test data are drawn from the same underlying distribution. However, this assumption is violated in almost all practical applications: machine learning systems are regularly tested under distribution shift, due to changing temporal correlations, atypical end users, or other factors. In this work, we consider the problem setting of domain generalization, where the training data are structured into domains and there may be multiple test time shifts, corresponding to new domains or domain distributions. Most prior methods aim to learn a single robust model or invariant feature space that performs well on all domains. In contrast, we aim to learn models that adapt at test time to domain shift using unlabeled test points. Our primary contribution is to introduce the framework of adaptive risk minimization (ARM), in which models are directly optimized for effective adaptation to shift by learning to adapt on the training domains. Compared to prior methods for robustness, invariance, and adaptation, ARM methods provide performance gains of 1-4% test accuracy on a number of image classification problems exhibiting domain shift.
****
Learning State Representations from Random Deep Action-conditional Predictions
Zeyu Zheng, Vivek Veeriah, Risto Vuorio, Richard L Lewis, Satinder Singh
Our main contribution in this work is an empirical finding that random General Value Functions (GVFs), i.e., deep action-conditional predictions---random both in what feature of observations they predict as well as in the sequence of actions the predictions are conditioned upon---form good auxiliary tasks for reinforcement learning (RL) problems. In particular, we show that random deep action-conditional predictions when used as auxiliary tasks yield state representations that produce control performance competitive with state-of-the-art hand-crafted auxiliary tasks like value prediction, pixel control, and CURL in both Atari and DeepMind Lab tasks. In another set of experiments we stop the gradients from the RL part of the network to the state representation learning part of the network and show, perhaps surprisingly, that the auxiliary tasks alone are sufficient to learn state representations good enough to outperform an end-to-end trained actor-critic baseline. We opensourced our code at https://github.com/Hwhitetooth/random_gvfs.
****
Mixability made efficient: Fast online multiclass logistic regression
Rémi Jézéquel, Pierre Gaillard, Alessandro Rudi
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-auth

ors prior to requesting a name change in the electronic proceedings.
****
Tracking People with 3D Representations
Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, Jitendra Malik
We present a novel approach for tracking multiple people in video. Unlike past approaches which employ 2D representations, we focus on using 3D representations of people, located in three-dimensional space. To this end, we develop a method, Human Mesh and Appearance Recovery (HMAR) which in addition to extracting the 3D geometry of the person as a SMPL mesh, also extracts appearance as a texture map on the triangles of the mesh. This serves as a 3D representation for appearance that is robust to viewpoint and pose changes. Given a video clip, we first detect bounding boxes corresponding to people, and for each one, we extract 3D appearance, pose, and location information using HMAR. These embedding vectors are then sent to a transformer, which performs spatio-temporal aggregation of the representations over the duration of the sequence. The similarity of the resulting representations is used to solve for associations that assigns each person to a tracklet. We evaluate our approach on the Posetrack, MuPoTs and AVA datasets. We find that 3D representations are more effective than 2D representations for tracking in these settings, and we obtain state-of-the-art performance. Code and results are available at: https://brjathu.github.io/T3DP.
****
Off-Policy Risk Assessment in Contextual Bandits
Audrey Huang, Liu Leqi, Zachary Lipton, Kamyar Azizzadenesheli
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 Denoising via GainTuning
Sreyas Mohan, Joshua L Vincent, Ramon Manzorro, Peter Crozier, Carlos Fernandez-Granda, Eero Simoncelli
Deep convolutional neural networks (CNNs) for image denoising are typically trained on large datasets. These models achieve the current state of the art, but they do not generalize well to data that deviate from the training distribution. Recent work has shown that it is possible to train denoisers on a single noisy image. These models adapt to the features of the test image, but their performance is limited by the small amount of information used to train them. Here we propose "GainTuning'', a methodology by which CNN models pre-trained on large datasets can be adaptively and selectively adjusted for individual test images. To avoid overfitting, GainTuning optimizes a single multiplicative scaling parameter (the “Gain”) of each channel in the convolutional layers of the CNN. We show that GainTuning improves state-of-the-art CNNs on standard image-denoising benchmarks, boosting their denoising performance on nearly every image in a held-out test set. These adaptive improvements are even more substantial for test images differing systematically from the training data, either in noise level or image type. We illustrate the potential of adaptive GainTuning in a scientific application to transmission-electron-microscope images, using a CNN that is pre-trained on synthetic data. In contrast to the existing methodology, GainTuning is able to faithfully reconstruct the structure of catalytic nanoparticles from these data at extremely low signal-to-noise ratios.
****
Optimal Sketching for Trace Estimation
Shuli Jiang, Hai Pham, David Woodruff, Richard Zhang
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Estimating Multi-cause Treatment Effects via Single-cause Perturbation
Zhaozhi Qian, Alicia Curth, Mihaela van der Schaar

Most existing methods for conditional average treatment effect estimation are de
signed to estimate the effect of a single cause - only one variable can be inter
vened on at one time. However, many applications involve simultaneous interventi
on on multiple variables, which leads to multi-cause treatment effect problems.
The multi-cause problem is challenging because one needs to overcome the confoun
ding bias for a large number of treatment groups, each with a different cause co
mbination. The combinatorial nature of the problem also leads to severe data sca
rcity - we only observe one factual outcome out of many potential outcomes. In t
his work, we propose Single-cause Perturbation (SCP), a novel two-step procedure
to estimate the multi-cause treatment effect. SCP starts by augmenting the obse
rvational dataset with the estimated potential outcomes under single-cause inter
ventions. It then performs covariate adjustment on the augmented dataset to obta
in the estimator. SCP is agnostic to the exact choice of algorithm in either ste
p. We show formally that the procedure is valid under standard assumptions in ca
usal inference. We demonstrate the performance gain of SCP on extensive syntheti
c and semi-synthetic experiments.
****
Be Confident! Towards Trustworthy Graph Neural Networks via Confidence Calibrati
on
Xiao Wang, Hongrui Liu, Chuan Shi, Cheng Yang
Despite Graph Neural Networks (GNNs) have achieved remarkable accuracy, whether
the results are trustworthy is still unexplored. Previous studies suggest that m
any modern neural networks are over-confident on the predictions, however, surpr
isingly, we discover that GNNs are primarily in the opposite direction, i.e., GN
Ns are under-confident. Therefore, the confidence calibration for GNNs is highly
desired. In this paper, we propose a novel trustworthy GNN model by designing a
topology-aware post-hoc calibration function. Specifically, we first verify tha
t the confidence distribution in a graph has homophily property, and this findin
g inspires us to design a calibration GNN model (CaGCN) to learn the calibration
function. CaGCN is able to obtain a unique transformation from logits of GNNs t
o the calibrated confidence for each node, meanwhile, such transformation is abl
e to preserve the order between classes, satisfying the accuracy-preserving prop
erty. Moreover, we apply the calibration GNN to self-training framework, showing
that more trustworthy pseudo labels can be obtained with the calibrated confide
nce and further improve the performance. Extensive experiments demonstrate the e
ffectiveness of our proposed model in terms of both calibration and accuracy.
****
Learning Riemannian metric for disease progression modeling
Samuel Gruffaz, Pierre-Emmanuel Poulet, Etienne Maheux, Bruno Jedynak, Stanley D
URRLEMAN
Linear mixed-effect models provide a natural baseline for estimating disease pro
gression using longitudinal data. They provide interpretable models at the cost
of modeling assumptions on the progression profiles and their variability across
subjects. A significant improvement is to embed the data in a Riemannian manifo
ld and learn patient-specific trajectories distributed around a central geodesic
. A few interpretable parameters characterize subject trajectories at the cost o
f a prior choice of the metric, which determines the shape of the trajectories.
We extend this approach by learning the metric from the data allowing more flexi
bility while keeping the interpretability. Specifically, we learn the metric as
the push-forward of the Euclidean metric by a diffeomorphism. This diffeomorphis
m is estimated iteratively as the composition of radial basis functions belongin
g to a reproducible kernel Hilbert space. The metric update allows us to improve
the forecasting of imaging and clinical biomarkers in the Alzheimer’s Disease N
euroimaging Initiative (ADNI) cohort. Our results compare favorably to the 56 me
thods benchmarked in the TADPOLE challenge.
****
Bias and variance of the Bayesian-mean decoder
Arthur Prat-Carrabin, Michael Woodford
Perception, in theoretical neuroscience, has been modeled as the encoding of ext
ernal stimuli into internal signals, which are then decoded. The Bayesian mean i

s an important decoder, as it is optimal for purposes of both estimation and discrimination. We present widely-applicable approximations to the bias and to the variance of the Bayesian mean, obtained under the minimal and biologically-relevant assumption that the encoding results from a series of independent, though not necessarily identically-distributed, signals. Simulations substantiate the accuracy of our approximations in the small-noise regime. The bias of the Bayesian mean comprises two components: one driven by the prior, and one driven by the precision of the encoding. If the encoding is 'efficient', the two components have opposite effects; their relative strengths are determined by the objective that the encoding optimizes. The experimental literature on perception reports both 'Bayesian' biases directed towards prior expectations, and opposite, 'anti-Bayesian' biases. We show that different tasks are indeed predicted to yield such contradictory biases, under a consistently-optimal encoding-decoding model. Moreover, we recover Wei and Stocker's "law of human perception", a relation between the bias of the Bayesian mean and the derivative of its variance, and show how the coefficient of proportionality in this law depends on the task at hand. Our results provide a parsimonious theory of optimal perception under constraints, in which encoding and decoding are adapted both to the prior and to the task faced by the observer.
****
MIRACLE: Causally-Aware Imputation via Learning Missing Data Mechanisms
Trent Kyono, Yao Zhang, Alexis Bellot, Mihaela van der Schaar
Missing data is an important problem in machine learning practice. Starting from the premise that imputation methods should preserve the causal structure of the data, we develop a regularization scheme that encourages any baseline imputation method to be causally consistent with the underlying data generating mechanism. Our proposal is a causally-aware imputation algorithm (MIRACLE). MIRACLE iteratively refines the imputation of a baseline by simultaneously modeling the missingness generating mechanism, encouraging imputation to be consistent with the causal structure of the data. We conduct extensive experiments on synthetic and a variety of publicly available datasets to show that MIRACLE is able to consistently improve imputation over a variety of benchmark methods across all three missingness scenarios: at random, completely at random, and not at random.
****
Efficient Training of Visual Transformers with Small Datasets
Yahui Liu, Enver Sangineto, Wei Bi, Nicu Sebe, Bruno Lepri, Marco Nadai
Visual Transformers (VTs) are emerging as an architectural paradigm alternative to Convolutional networks (CNNs). Differently from CNNs, VTs can capture global relations between image elements and they potentially have a larger representation capacity. However, the lack of the typical convolutional inductive bias makes these models more data hungry than common CNNs. In fact, some local properties of the visual domain which are embedded in the CNN architectural design, in VTs should be learned from samples. In this paper, we empirically analyse different VTs, comparing their robustness in a small training set regime, and we show that, despite having a comparable accuracy when trained on ImageNet, their performance on smaller datasets can be largely different. Moreover, we propose an auxiliary self-supervised task which can extract additional information from images with only a negligible computational overhead. This task encourages the VTs to learn spatial relations within an image and makes the VT training much more robust when training data is scarce. Our task is used jointly with the standard (supervised) training and it does not depend on specific architectural choices, thus it can be easily plugged in the existing VTs. Using an extensive evaluation with different VTs and datasets, we show that our method can improve (sometimes dramatically) the final accuracy of the VTs. Our code is available at: https://github.com/yhlleo/VTs-Drloc.
****
Small random initialization is akin to spectral learning: Optimization and generalization guarantees for overparameterized low-rank matrix reconstruction
Dominik Stöger, Mahdi Soltanolkotabi
Recently there has been significant theoretical progress on understanding the co

nvergence and generalization of gradient-based methods on nonconvex losses with overparameterized models. Nevertheless, many aspects of optimization and generalization and in particular the critical role of small random initialization are not fully understood. In this paper, we take a step towards demystifying this role by proving that small random initialization followed by a few iterations of gradient descent behaves akin to popular spectral methods. We also show that this implicit spectral bias from small random initialization, which is provably more prominent for overparameterized models, also puts the gradient descent iterations on a particular trajectory towards solutions that are not only globally optimal but also generalize well. Concretely, we focus on the problem of reconstructing a low-rank matrix from a few measurements via a natural nonconvex formulation. In this setting, we show that the trajectory of the gradient descent iterations from small random initialization can be approximately decomposed into three phases: (I) a spectral or alignment phase where we show that that the iterates have an implicit spectral bias akin to spectral initialization allowing us to show that at the end of this phase the column space of the iterates and the underlying low-rank matrix are sufficiently aligned, (II) a saddle avoidance/refinement phase where we show that the trajectory of the gradient iterates moves away from certain degenerate saddle points, and (III) a local refinement phase where we show that after avoiding the saddles the iterates converge quickly to the underlying low-rank matrix. Underlying our analysis are insights for the analysis of overparameterized nonconvex optimization schemes that may have implications for computational problems beyond low-rank reconstruction.
****
Efficient Combination of Rematerialization and Offloading for Training DNNs
Olivier Beaumont, Lionel Eyraud-Dubois, Alena Shilova
Rematerialization and offloading are two well known strategies to save memory during the training phase of deep neural networks, allowing data scientists to consider larger models, batch sizes or higher resolution data. Rematerialization trades memory for computation time, whereas Offloading trades memory for data movements. As these two resources are independent, it is appealing to consider the simultaneous combination of both strategies to save even more memory. We precisely model the costs and constraints corresponding to Deep Learning frameworks such as PyTorch or Tensorflow, we propose optimal algorithms to find a valid sequence of memory-constrained operations and finally, we evaluate the performance of proposed algorithms on realistic networks and computation platforms. Our experiments show that the possibility to offload can remove one third of the overhead of rematerialization, and that together they can reduce the memory used for activations by a factor 4 to 6, with an overhead below 20%.
****
Particle Cloud Generation with Message Passing Generative Adversarial Networks
Raghav Kansal, Javier Duarte, Hao Su, Breno Orzari, Thiago Tomei, Maurizio Pierini, Mary Touranakou, jean-roch vlimant, Dimitrios Gunopulos
In high energy physics (HEP), jets are collections of correlated particles produced ubiquitously in particle collisions such as those at the CERN Large Hadron Collider (LHC). Machine learning (ML)-based generative models, such as generative adversarial networks (GANs), have the potential to significantly accelerate LHC jet simulations. However, despite jets having a natural representation as a set of particles in momentum-space, a.k.a. a particle cloud, there exist no generative models applied to such a dataset. In this work, we introduce a new particle cloud dataset (JetNet), and apply to it existing point cloud GANs. Results are evaluated using (1) 1-Wasserstein distances between high- and low-level feature distributions, (2) a newly developed Fréchet ParticleNet Distance, and (3) the coverage and (4) minimum matching distance metrics. Existing GANs are found to be inadequate for physics applications, hence we develop a new message passing GAN (MPGAN), which outperforms existing point cloud GANs on virtually every metric and shows promise for use in HEP. We propose JetNet as a novel point-cloud-style dataset for the ML community to experiment with, and set MPGAN as a benchmark to improve upon for future generative models. Additionally, to facilitate research and improve accessibility and reproducibility in this area, we release the open

-source JetNet Python package with interfaces for particle cloud datasets, imple mentations for evaluation and loss metrics, and more tools for ML in HEP develop ment.
****
CoFiNet: Reliable Coarse-to-fine Correspondences for Robust PointCloud Registrat ion
Hao Yu, Fu Li, Mahdi Saleh, Benjamin Busam, Slobodan Ilic
We study the problem of extracting correspondences between a pair of point cloud s for registration. For correspondence retrieval, existing works benefit from ma tching sparse keypoints detected from dense points but usually struggle to guara ntee their repeatability. To address this issue, we present CoFiNet - Coarse-to-Fine Network which extracts hierarchical correspondences from coarse to fine wit hout keypoint detection. On a coarse scale and guided by a weighting scheme, our model firstly learns to match down-sampled nodes whose vicinity points share mo re overlap, which significantly shrinks the search space of a consecutive stage. On a finer scale, node proposals are consecutively expanded to patches that con sist of groups of points together with associated descriptors. Point corresponde nces are then refined from the overlap areas of corresponding patches, by a dens ity-adaptive matching module capable to deal with varying point density. Extensi ve evaluation of CoFiNet on both indoor and outdoor standard benchmarks shows ou r superiority over existing methods. Especially on 3DLoMatch where point clouds share less overlap, CoFiNet significantly outperforms state-of-the-art approache s by at least 5% on Registration Recall, with at most two-third of their paramet ers.
****
Partial success in closing the gap between human and machine vision
Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Mat thias Bethge, Felix A. Wichmann, Wieland Brendel
A few years ago, the first CNN surpassed human performance on ImageNet. However, it soon became clear that machines lack robustness on more challenging test cas es, a major obstacle towards deploying machines "in the wild" and towards obtain ing better computational models of human visual perception. Here we ask: Are we making progress in closing the gap between human and machine vision? To answer t his question, we tested human observers on a broad range of out-of-distribution (OOD) datasets, recording 85,120 psychophysical trials across 90 participants. W e then investigated a range of promising machine learning developments that cruc ially deviate from standard supervised CNNs along three axes: objective function (self-supervised, adversarially trained, CLIP language-image training), archite cture (e.g. vision transformers), and dataset size (ranging from 1M to 1B).Our f indings are threefold. (1.) The longstanding distortion robustness gap between h umans and CNNs is closing, with the best models now exceeding human feedforward performance on most of the investigated OOD datasets. (2.) There is still a subs tantial image-level consistency gap, meaning that humans make different errors t han models. In contrast, most models systematically agree in their categorisatio n errors, even substantially different ones like contrastive self-supervised vs. standard supervised models. (3.) In many cases, human-to-model consistency impr oves when training dataset size is increased by one to three orders of magnitude . Our results give reason for cautious optimism: While there is still much room for improvement, the behavioural difference between human and machine vision is narrowing. In order to measure future progress, 17 OOD datasets with image-level human behavioural data and evaluation code are provided as a toolbox and benchm ark at: https://github.com/bethgelab/model-vs-human/
****
LLC: Accurate, Multi-purpose Learnt Low-dimensional Binary Codes
Aditya Kusupati, Matthew Wallingford, Vivek Ramanujan, Raghav Somani, Jae Sung P ark, Krishna Pillutla, Prateek Jain, Sham Kakade, Ali Farhadi
Requests for name changes in the electronic proceedings will be accepted with 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.

****
Analytic Insights into Structure and Rank of Neural Network Hessian Maps
Sidak Pal Singh, Gregor Bachmann, Thomas Hofmann
The Hessian of a neural network captures parameter interactions through second-o
rder derivatives of the loss. It is a fundamental object of study, closely tied
to various problems in deep learning, including model design, optimization, and
generalization. Most prior work has been empirical, typically focusing on low-ra
nk approximations and heuristics that are blind to the network structure. In co
ntrast, we develop theoretical tools to analyze the range of the Hessian map, wh
ich provide us with a precise understanding of its rank deficiency and the struc
tural reasons behind it. This yields exact formulas and tight upper bounds for t
he Hessian rank of deep linear networks --- allowing for an elegant interpretati
on in terms of rank deficiency. Moreover, we demonstrate that our bounds remain
faithful as an estimate of the numerical Hessian rank, for a larger class of mod
els such as rectified and hyperbolic tangent networks. Further, we also investig
ate the implications of model architecture (e.g.~width, depth, bias) on the rank
 deficiency. Overall, our work provides novel insights into the source and exten
t of redundancy in overparameterized neural networks.
****
Well-tuned Simple Nets Excel on Tabular Datasets
Arlind Kadra, Marius Lindauer, Frank Hutter, Josif Grabocka
Tabular datasets are the last "unconquered castle" for deep learning, with tradi
tional ML methods like Gradient-Boosted Decision Trees still performing strongly
 even against recent specialized neural architectures. In this paper, we hypothe
size that the key to boosting the performance of neural networks lies in rethink
ing the joint and simultaneous application of a large set of modern regularizati
on techniques. As a result, we propose regularizing plain Multilayer Perceptron
(MLP) networks by searching for the optimal combination/cocktail of 13 regulariz
ation techniques for each dataset using a joint optimization over the decision o
n which regularizers to apply and their subsidiary hyperparameters. We empirical
ly assess the impact of these regularization cocktails for MLPs in a large-scale
 empirical study comprising 40 tabular datasets and demonstrate that (i) well-re
gularized plain MLPs significantly outperform recent state-of-the-art specialize
d neural network architectures, and (ii) they even outperform strong traditional
 ML methods, such as XGBoost.
****
POODLE: Improving Few-shot Learning via Penalizing Out-of-Distribution Samples
Duong Le, Khoi Duc Nguyen, Khoi Nguyen, Quoc-Huy Tran, Rang Nguyen, Binh-Son Hua
In this work, we propose to use out-of-distribution samples, i.e., unlabeled sam
ples coming from outside the target classes, to improve few-shot learning. Speci
fically, we exploit the easily available out-of-distribution samples to drive th
e classifier to avoid irrelevant features by maximizing the distance from protot
ypes to out-of-distribution samples while minimizing that of in-distribution sam
ples (i.e., support, query data). Our approach is simple to implement, agnostic
to feature extractors, lightweight without any additional cost for pre-training,
 and applicable to both inductive and transductive settings. Extensive experimen
ts on various standard benchmarks demonstrate that the proposed method consisten
tly improves the performance of pretrained networks with different architectures
.
****
Combinatorial Pure Exploration with Bottleneck Reward Function
Yihan Du, Yuko Kuroki, 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.
****
Densely connected normalizing flows
Matej Grci■, Ivan Grubiši■, Siniša Šegvi■
Normalizing flows are bijective mappings between inputs and latent representatio

ns with a fully factorized distribution. They are very attractive due to exact l
ikelihood evaluation and efficient sampling. However, their effective capacity i
s often insufficient since the bijectivity constraint limits the model width. We
 address this issue by incrementally padding intermediate representations with n
oise. We precondition the noise in accordance with previous invertible units, wh
ich we describe as cross-unit coupling. Our invertible glow-like modules increas
e the model expressivity by fusing a densely connected block with Nyström self-a
ttention. We refer to our architecture as DenseFlow since both cross-unit and in
tra-module couplings rely on dense connectivity. Experiments show significant im
provements due to the proposed contributions and reveal state-of-the-art density
 estimation under moderate computing budgets.
****
Snowflake: Scaling GNNs to high-dimensional continuous control via parameter fre
ezing
Charles Blake, Vitaly Kurin, Maximilian Igl, Shimon Whiteson
Recent research has shown that graph neural networks (GNNs) can learn policies f
or locomotion control that are as effective as a typical multi-layer perceptron
(MLP), with superior transfer and multi-task performance. However, results have
so far been limited to training on small agents, with the performance of GNNs de
teriorating rapidly as the number of sensors and actuators grows. A key motivati
on for the use of GNNs in the supervised learning setting is their applicability
 to large graphs, but this benefit has not yet been realised for locomotion cont
rol. We show that poor scaling in GNNs is a result of increasingly unstable poli
cy updates, caused by overfitting in parts of the network during training. To co
mbat this, we introduce Snowflake, a GNN training method for high-dimensional co
ntinuous control that freezes parameters in selected parts of the network. Snowf
lake significantly boosts the performance of GNNs for locomotion control on larg
e agents, now matching the performance of MLPs while offering superior transfer
properties.
****
Subgame solving without common knowledge
Brian Zhang, Tuomas Sandholm
In imperfect-information games, subgame solving is significantly more challengin
g than in perfect-information games, but in the last few years, such techniques
have been developed. They were the key ingredient to the milestone of superhuman
 play in no-limit Texas hold'em poker. Current subgame-solving techniques analyz
e the entire common-knowledge closure of the player's current information set, t
hat is, the smallest set of nodes within which it is common knowledge that the c
urrent node lies. While this is acceptable in games like poker where the common-
knowledge closure is relatively small, many practical games have more complex in
formation structure, which renders the common-knowledge closure impractically la
rge to enumerate or even reasonably approximate. We introduce an approach that o
vercomes this obstacle, by instead working with only low-order knowledge. Our ap
proach allows an agent, upon arriving at an infoset, to basically prune any node
 that is no longer reachable, thereby massively reducing the game tree size rela
tive to the common-knowledge subgame. We prove that, as is, our approach can inc
rease exploitability compared to the blueprint strategy. However, we develop thr
ee avenues by which safety can be guaranteed. First, safety is guaranteed if the
 results of subgame solves are incorporated back into the blueprint. Second, we
provide a method where safety is achieved by limiting the infosets at which subg
ame solving is performed. Third, we prove that our approach, when applied at eve
ry infoset reached during play, achieves a weaker notion of equilibrium, which w
e coin affine equilibrium, and which may be of independent interest. We show tha
t affine equilibria cannot be exploited by any Nash strategy of the opponent, so
 an opponent who wishes to exploit must open herself to counter-exploitation. Ev
en without the safety-guaranteeing additions, experiments on medium-sized games
show that our approach always reduced exploitability in practical games even whe
n applied at every infoset, and a depth-limited version of it led to---to our kn
owledge---the first strong AI for the challenge problem dark chess.
****

Fair Algorithms for Multi-Agent Multi-Armed Bandits
Safwan Hossain, Evi Micha, Nisarg 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.
****
VAST: Value Function Factorization with Variable Agent Sub-Teams
Thomy Phan, Fabian Ritz, Lenz Belzner, Philipp Altmann, Thomas Gabor, Claudia Linnhoff-Popien
Value function factorization (VFF) is a popular approach to cooperative multi-agent reinforcement learning in order to learn local value functions from global rewards. However, state-of-the-art VFF is limited to a handful of agents in most domains. We hypothesize that this is due to the flat factorization scheme, where the VFF operator becomes a performance bottleneck with an increasing number of agents. Therefore, we propose VFF with variable agent sub-teams (VAST). VAST approximates a factorization for sub-teams which can be defined in an arbitrary way and vary over time, e.g., to adapt to different situations. The sub-team values are then linearly decomposed for all sub-team members. Thus, VAST can learn on a more focused and compact input representation of the original VFF operator. We evaluate VAST in three multi-agent domains and show that VAST can significantly outperform state-of-the-art VFF, when the number of agents is sufficiently large.
****
On the Stochastic Stability of Deep Markov Models
Jan Drgona, Sayak Mukherjee, Jiaxin Zhang, Frank Liu, Mahantesh Halappanavar
Deep Markov models (DMM) are generative models which are scalable and expressive generalization of Markov models for representation, learning, and inference problems. However, the fundamental stochastic stability guarantees of such models have not been thoroughly investigated. In this paper, we present a novel stability analysis method and provide sufficient conditions of DMM's stochastic stability. The proposed stability analysis is based on the contraction of probabilistic maps modeled by deep neural networks. We make connections between the spectral properties of neural network's weights and different types of used activation function on the stability and overall dynamic behavior of DMMs with Gaussian distributions. Based on the theory, we propose a few practical methods for designing constrained DMMs with guaranteed stability. We empirically substantiate our theoretical results via intuitive numerical experiments using the proposed stability constraints.
****
Multiwavelet-based Operator Learning for Differential Equations
Gaurav Gupta, Xiongye Xiao, Paul Bogdan
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Intermediate Layers Matter in Momentum Contrastive Self Supervised Learning
Aakash Kaku, Sahana Upadhya, Narges Razavian
We show that bringing intermediate layers' representations of two augmented versions of an image closer together in self-supervised learning helps to improve the momentum contrastive (MoCo) method. To this end, in addition to the contrastive loss, we minimize the mean squared error between the intermediate layer representations or make their cross-correlation matrix closer to an identity matrix. Both loss objectives either outperform standard MoCo, or achieve similar performances on three diverse medical imaging datasets: NIH-Chest Xrays, Breast Cancer Histopathology, and Diabetic Retinopathy. The gains of the improved MoCo are especially large in a low-labeled data regime (e.g. 1% labeled data) with an average gain of 5% across three datasets. We analyze the models trained using our novel approach via feature similarity analysis and layer-wise probing. Our analysis r

eveals that models trained via our approach have higher feature reuse compared to a standard MoCo and learn informative features earlier in the network. Finally, by comparing the output probability distribution of models fine-tuned on small versus large labeled data, we conclude that our proposed method of pre-training leads to lower Kolmogorov–Smirnov distance, as compared to a standard MoCo. This provides additional evidence that our proposed method learns more informative features in the pre-training phase which could be leveraged in a low-labeled data regime.
****
An Efficient Pessimistic-Optimistic Algorithm for Stochastic Linear Bandits with General Constraints
Xin Liu, Bin Li, Pengyi Shi, Lei Ying
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Efficiently Learning One Hidden Layer ReLU Networks From Queries
Sitan Chen, Adam Klivans, Raghu Meka
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 Nonparametric Volterra Kernels with Gaussian Processes
Magnus Ross, Michael T Smith, Mauricio Álvarez
This paper introduces a method for the nonparametric Bayesian learning of nonlinear operators, through the use of the Volterra series with kernels represented using Gaussian processes (GPs), which we term the nonparametric Volterra kernels model (NVKM). When the input function to the operator is unobserved and has a GP prior, the NVKM constitutes a powerful method for both single and multiple output regression, and can be viewed as a nonlinear and nonparametric latent force model. When the input function is observed, the NVKM can be used to perform Bayesian system identification. We use recent advances in efficient sampling of explicit functions from GPs to map process realisations through the Volterra series without resorting to numerical integration, allowing scalability through doubly stochastic variational inference, and avoiding the need for Gaussian approximations of the output processes. We demonstrate the performance of the model for both multiple output regression and system identification using standard benchmarks.
****
DiBS: Differentiable Bayesian Structure Learning
Lars Lorch, Jonas Rothfuss, Bernhard Schölkopf, Andreas Krause
Bayesian structure learning allows inferring Bayesian network structure from data while reasoning about the epistemic uncertainty---a key element towards enabling active causal discovery and designing interventions in real world systems. In this work, we propose a general, fully differentiable framework for Bayesian structure learning (DiBS) that operates in the continuous space of a latent probabilistic graph representation. Contrary to existing work, DiBS is agnostic to the form of the local conditional distributions and allows for joint posterior inference of both the graph structure and the conditional distribution parameters. This makes our formulation directly applicable to posterior inference of nonstandard Bayesian network models, e.g., with nonlinear dependencies encoded by neural networks. Using DiBS, we devise an efficient, general purpose variational inference method for approximating distributions over structural models. In evaluations on simulated and real-world data, our method significantly outperforms related approaches to joint posterior inference.
****
Nonparametric estimation of continuous DPPs with kernel methods
Michaël Fanuel, Rémi Bardenet
Determinantal Point Process (DPPs) are statistical models for repulsive point pa

tterns. Both sampling and inference are tractable for DPPs, a rare feature among models with negative dependence that explains their popularity in machine learning and spatial statistics. Parametric and nonparametric inference methods have been proposed in the finite case, i.e. when the point patterns live in a finite ground set. In the continuous case, only parametric methods have been investigated, while nonparametric maximum likelihood for DPPs -- an optimization problem over trace-class operators -- has remained an open question. In this paper, we show that a restricted version of this maximum likelihood (MLE) problem falls within the scope of a recent representer theorem for nonnegative functions in an RKHS. This leads to a finite-dimensional problem, with strong statistical ties to the original MLE. Moreover, we propose, analyze, and demonstrate a fixed point algorithm to solve this finite-dimensional problem. Finally, we also provide a controlled estimate of the correlation kernel of the DPP, thus providing more interpretability.
****
FINE Samples for Learning with Noisy Labels
Taehyeon Kim, Jongwoo Ko, sangwook Cho, JinHwan Choi, Se-Young Yun
Modern deep neural networks (DNNs) become frail when the datasets contain noisy (incorrect) class labels. Robust techniques in the presence of noisy labels can be categorized into two folds: developing noise-robust functions or using noise-cleansing methods by detecting the noisy data. Recently, noise-cleansing methods have been considered as the most competitive noisy-label learning algorithms. Despite their success, their noisy label detectors are often based on heuristics more than a theory, requiring a robust classifier to predict the noisy data with loss values. In this paper, we propose a novel detector for filtering label noise. Unlike most existing methods, we focus on each data's latent representation dynamics and measure the alignment between the latent distribution and each representation using the eigen decomposition of the data gram matrix. Our framework, coined as filtering noisy instances via their eigenvectors (FINE), provides a robust detector with derivative-free simple methods having theoretical guarantees. Under our framework, we propose three applications of the FINE: sample-selection approach, semi-supervised learning approach, and collaboration with noise-robust loss functions. Experimental results show that the proposed methods consistently outperform corresponding baselines for all three applications on various benchmark datasets.
****
Residual2Vec: Debiasing graph embedding with random graphs
Sadamori Kojaku, Jisung Yoon, Isabel Constantino, Yong-Yeol Ahn
Graph embedding maps a graph into a convenient vector-space representation for graph analysis and machine learning applications. Many graph embedding methods hinge on a sampling of context nodes based on random walks. However, random walks can be a biased sampler due to the structural properties of graphs. Most notably, random walks are biased by the degree of each node, where a node is sampled proportionally to its degree. The implication of such biases has not been clear, particularly in the context of graph representation learning. Here, we investigate the impact of the random walks' bias on graph embedding and propose residual2vec, a general graph embedding method that can debias various structural biases in graphs by using random graphs. We demonstrate that this debiasing not only improves link prediction and clustering performance but also allows us to explicitly model salient structural properties in graph embedding.
****
Benign Overfitting in Multiclass Classification: All Roads Lead to Interpolation
Ke Wang, Vidya Muthukumar, Christos Thrampoulidis
The growing literature on "benign overfitting" in overparameterized models has been mostly restricted to regression or binary classification settings; however, most success stories of modern machine learning have been recorded in multiclass settings. Motivated by this discrepancy, we study benign overfitting in multiclass linear classification. Specifically, we consider the following popular training algorithms on separable data: (i) empirical risk minimization (ERM) with cross-entropy loss, which converges to the multiclass support vector machine (SVM)

solution; (ii) ERM with least-squares loss, which converges to the min-norm inte rpolating (MNI) solution; and, (iii) the one-vs-all SVM classifier. Our first ke y finding is that under a simple sufficient condition, all three algorithms lead to classifiers that interpolate the training data and have equal accuracy. When the data is generated from Gaussian mixtures or a multinomial logistic model, t his condition holds under high enough effective overparameterization. Second, we derive novel error bounds on the accuracy of the MNI classifier, thereby showin g that all three training algorithms lead to benign overfitting under sufficient overparameterization. Ultimately, our analysis shows that good generalization i s possible for SVM solutions beyond the realm in which typical margin-based boun ds apply.

****

Instance-Dependent Bounds for Zeroth-order Lipschitz Optimization with Error Cer tificates

Francois Bachoc, Tom Cesari, Sébastien Gerchinovitz

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

****

Training Neural Networks with Fixed Sparse Masks

Yi-Lin Sung, Varun Nair, Colin A. Raffel

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

****

VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text

Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, Boqing Gong

We present a framework for learning multimodal representations from unlabeled da ta using convolution-free Transformer architectures. Specifically, our Video-Aud io-Text Transformer (VATT) takes raw signals as inputs and extracts multimodal r epresentations that are rich enough to benefit a variety of downstream tasks. We train VATT end-to-end from scratch using multimodal contrastive losses and eval uate its performance by the downstream tasks of video action recognition, audio event classification, image classification, and text-to-video retrieval. Further more, we study a modality-agnostic single-backbone Transformer by sharing weight s among the three modalities. We show that the convolution-free VATT outperforms state-of-the-art ConvNet-based architectures in the downstream tasks. Especiall y, VATT's vision Transformer achieves the top-1 accuracy of 82.1% on Kinetics-40 0, 83.6% on Kinetics-600, 72.7% on Kinetics-700, and 41.1% on Moments in Time, n ew records while avoiding supervised pre-training. Transferring to image classif ication leads to 78.7% top-1 accuracy on ImageNet compared to 64.7% by training the same Transformer from scratch, showing the generalizability of our model des pite the domain gap between videos and images. VATT's audio Transformer also set s a new record on waveform-based audio event recognition by achieving the mAP of 39.4% on AudioSet without any supervised pre-training.

****

Analyzing the Generalization Capability of SGLD Using Properties of Gaussian Cha nnels

Hao Wang, Yizhe Huang, Rui Gao, Flavio Calmon

Optimization is a key component for training machine learning models and has a s trong impact on their generalization. In this paper, we consider a particular op timization method---the stochastic gradient Langevin dynamics (SGLD) algorithm---and investigate the generalization of models trained by SGLD. We derive a new g eneralization bound by connecting SGLD with Gaussian channels found in informati on and communication theory. Our bound can be computed from the training data an d incorporates the variance of gradients for quantifying a particular kind of "s

harpness" of the loss landscape. We also consider a closely related algorithm wi
th SGLD, namely differentially private SGD (DP-SGD). We prove that the generaliz
ation capability of DP-SGD can be amplified by iteration. Specifically, our boun
d can be sharpened by including a time-decaying factor if the DP-SGD algorithm o
utputs the last iterate while keeping other iterates hidden. This decay factor e
nables the contribution of early iterations to our bound to reduce with time and
 is established by strong data processing inequalities---a fundamental tool in i
nformation theory. We demonstrate our bound through numerical experiments, showi
ng that it can predict the behavior of the true generalization gap.
****
Learning to Schedule Heuristics in Branch and Bound
Antonia Chmiela, Elias Khalil, Ambros Gleixner, Andrea Lodi, Sebastian Pokutta
Primal heuristics play a crucial role in exact solvers for Mixed Integer Program
ming (MIP). While solvers are guaranteed to find optimal solutions given suffici
ent time, real-world applications typically require finding good solutions early
 on in the search to enable fast decision-making. While much of MIP research foc
uses on designing effective heuristics, the question of how to manage multiple M
IP heuristics in a solver has not received equal attention. Generally, solvers f
ollow hard-coded rules derived from empirical testing on broad sets of instances
. Since the performance of heuristics is problem-dependent, using these general
rules for a particular problem might not yield the best performance. In this wor
k, we propose the first data-driven framework for scheduling heuristics in an ex
act MIP solver. By learning from data describing the performance of primal heuri
stics, we obtain a problem-specific schedule of heuristics that collectively fin
d many solutions at minimal cost. We formalize the learning task and propose an
efficient algorithm for computing such a schedule. Compared to the default setti
ngs of a state-of-the-art academic MIP solver, we are able to reduce the average
 primal integral by up to 49% on two classes of challenging instances.
****
On Training Implicit Models
Zhengyang Geng, Xin-Yu Zhang, Shaojie Bai, Yisen Wang, Zhouchen 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.
****
MLP-Mixer: An all-MLP Architecture for Vision
Ilya O. Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zha
i, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszk
oreit, Mario Lucic, Alexey Dosovitskiy
Convolutional Neural Networks (CNNs) are the go-to model for computer vision. Re
cently, attention-based networks, such as the Vision Transformer, have also beco
me popular. In this paper we show that while convolutions and attention are both
 sufficient for good performance, neither of them are necessary. We present MLP-
Mixer, an architecture based exclusively on multi-layer perceptrons (MLPs). MLP-
Mixer contains two types of layers: one with MLPs applied independently to image
 patches (i.e. "mixing" the per-location features), and one with MLPs applied ac
ross patches (i.e. "mixing" spatial information). When trained on large datasets
, or with modern regularization schemes, MLP-Mixer attains competitive scores on
 image classification benchmarks, with pre-training and inference cost comparabl
e to state-of-the-art models. We hope that these results spark further research
beyond the realms of well established CNNs and Transformers.
****
A Framework to Learn with Interpretation
Jayneel Parekh, Pavlo Mozharovskyi, Florence d'Alché-Buc
To tackle interpretability in deep learning, we present a novel framework to joi
ntly learn a predictive model and its associated interpretation model. The inter
preter provides both local and global interpretability about the predictive mode
l in terms of human-understandable high level attribute functions, with minimal
loss of accuracy. This is achieved by a dedicated architecture and well chosen r

egularization penalties. We seek for a small-size dictionary of high level attri
bute functions that take as inputs the outputs of selected hidden layers and who
se outputs feed a linear classifier. We impose strong conciseness on the activat
ion of attributes with an entropy-based criterion while enforcing fidelity to bo
th inputs and outputs of the predictive model. A detailed pipeline to visualize
the learnt features is also developed. Moreover, besides generating interpretabl
e models by design, our approach can be specialized to provide post-hoc interpre
tations for a pre-trained neural network. We validate our approach against sever
al state-of-the-art methods on multiple datasets and show its efficacy on both k
inds of tasks.
****

One Loss for All: Deep Hashing with a Single Cosine Similarity based Learning Ob
jective
Jiun Tian Hoe, Kam Woh Ng, Tianyu Zhang, Chee Seng Chan, Yi-Zhe Song, Tao Xiang
Requests for name changes in the electronic proceedings will be accepted with 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.
****

Fast and accurate randomized algorithms for low-rank tensor decompositions
Linjian Ma, Edgar Solomonik
Requests for name changes in the electronic proceedings will be accepted with 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.
****

Communication-efficient SGD: From Local SGD to One-Shot Averaging
Artin Spiridonoff, Alex Olshevsky, Yannis Paschalidis
Requests for name changes in the electronic proceedings will be accepted with 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.
****

Memory Efficient Meta-Learning with Large Images
John Bronskill, Daniela Massiceti, Massimiliano Patacchiola, Katja Hofmann, Seba
stian Nowozin, Richard Turner
Meta learning approaches to few-shot classification are computationally efficien
t at test time, requiring just a few optimization steps or single forward pass t
o learn a new task, but they remain highly memory-intensive to train. This limit
ation arises because a task's entire support set, which can contain up to 1000 i
mages, must be processed before an optimization step can be taken. Harnessing th
e performance gains offered by large images thus requires either parallelizing t
he meta-learner across multiple GPUs, which may not be available, or trade-offs
between task and image size when memory constraints apply. We improve on both op
tions by proposing LITE, a general and memory efficient episodic training scheme
 that enables meta-training on large tasks composed of large images on a single
GPU. We achieve this by observing that the gradients for a task can be decompose
d into a sum of gradients over the task's training images. This enables us to pe
rform a forward pass on a task's entire training set but realize significant mem
ory savings by back-propagating only a random subset of these images which we sh
ow is an unbiased approximation of the full gradient. We use LITE to train meta-
learners and demonstrate new state-of-the-art accuracy on the real-world ORBIT b
enchmark and 3 of the 4 parts of the challenging VTAB+MD benchmark relative to l
eading meta-learners. LITE also enables meta-learners to be competitive with tra
nsfer learning approaches but at a fraction of the test-time computational cost,
 thus serving as a counterpoint to the recent narrative that transfer learning i
s all you need for few-shot classification.
****

On the Power of Differentiable Learning versus PAC and SQ Learning
Emmanuel Abbe, Pritish Kamath, Eran Malach, Colin Sandon, Nathan Srebro

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

Can we globally optimize cross-validation loss? Quasiconvexity in ridge regression

Will Stephenson, Zachary Frangella, Madeleine Udell, Tamara Broderick

Models like LASSO and ridge regression are extensively used in practice due to their interpretability, ease of use, and strong theoretical guarantees. Cross-validation (CV) is widely used for hyperparameter tuning in these models, but do practical methods minimize the true out-of-sample loss? A recent line of research promises to show that the optimum of the CV loss matches the optimum of the out-of-sample loss (possibly after simple corrections). It remains to show how tractable it is to minimize the CV loss.In the present paper, we show that, in the case of ridge regression, the CV loss may fail to be quasiconvex and thus may have multiple local optima. We can guarantee that the CV loss is quasiconvex in at least one case: when the spectrum of the covariate matrix is nearly flat and the noise in the observed responses is not too high. More generally, we show that quasiconvexity status is independent of many properties of the observed data (response norm, covariate-matrix right singular vectors and singular-value scaling) and has a complex dependence on the few that remain. We empirically confirm our theory using simulated experiments.
****

Adaptive Proximal Gradient Methods for Structured Neural Networks

Jihun Yun, Aurelie C. Lozano, Eunho Yang

We consider the training of structured neural networks where the regularizer can be non-smooth and possibly non-convex. While popular machine learning libraries have resorted to stochastic (adaptive) subgradient approaches, the use of proximal gradient methods in the stochastic setting has been little explored and warrants further study, in particular regarding the incorporation of adaptivity. Towards this goal, we present a general framework of stochastic proximal gradient descent methods that allows for arbitrary positive preconditioners and lower semi-continuous regularizers. We derive two important instances of our framework: (i) the first proximal version of \textsc{Adam}, one of the most popular adaptive SGD algorithm, and (ii) a revised version of ProxQuant for quantization-specific regularizers, which improves upon the original approach by incorporating the effect of preconditioners in the proximal mapping computations. We provide convergence guarantees for our framework and show that adaptive gradient methods can have faster convergence in terms of constant than vanilla SGD for sparse data. Lastly, we demonstrate the superiority of stochastic proximal methods compared to subgradient-based approaches via extensive experiments. Interestingly, our results indicate that the benefit of proximal approaches over sub-gradient counterparts is more pronounced for non-convex regularizers than for convex ones.
****

Discovering and Achieving Goals via World Models

Russell Mendonca, Oleh Rybkin, Kostas Daniilidis, Danijar Hafner, Deepak Pathak

How can artificial agents learn to solve many diverse tasks in complex visual environments without any supervision? We decompose this question into two challenges: discovering new goals and learning to reliably achieve them. Our proposed agent, Latent Explorer Achiever (LEXA), addresses both challenges by learning a world model from image inputs and using it to train an explorer and an achiever policy via imagined rollouts. Unlike prior methods that explore by reaching previously visited states, the explorer plans to discover unseen surprising states through foresight, which are then used as diverse targets for the achiever to practice. After the unsupervised phase, LEXA solves tasks specified as goal images zero-shot without any additional learning. LEXA substantially outperforms previous approaches to unsupervised goal reaching, both on prior benchmarks and on a new challenging benchmark with 40 test tasks spanning across four robotic manipulation and locomotion domains. LEXA further achieves goals that require interacting

with multiple objects in sequence. Project page: https://orybkin.github.io/lexa/
****
Understanding and Improving Early Stopping for Learning with Noisy Labels
Yingbin Bai, Erkun Yang, Bo Han, Yanhua Yang, Jiatong Li, Yinian Mao, Gang Niu, Tongliang Liu
The memorization effect of deep neural network (DNN) plays a pivotal role in many state-of-the-art label-noise learning methods. To exploit this property, the early stopping trick, which stops the optimization at the early stage of training, is usually adopted. Current methods generally decide the early stopping point by considering a DNN as a whole. However, a DNN can be considered as a composition of a series of layers, and we find that the latter layers in a DNN are much more sensitive to label noise, while their former counterparts are quite robust. Therefore, selecting a stopping point for the whole network may make different DNN layers antagonistically affect each other, thus degrading the final performance. In this paper, we propose to separate a DNN into different parts and progressively train them to address this problem. Instead of the early stopping which trains a whole DNN all at once, we initially train former DNN layers by optimizing the DNN with a relatively large number of epochs. During training, we progressively train the latter DNN layers by using a smaller number of epochs with the preceding layers fixed to counteract the impact of noisy labels. We term the proposed method as progressive early stopping (PES). Despite its simplicity, compared with the traditional early stopping, PES can help to obtain more promising and stable results. Furthermore, by combining PES with existing approaches on noisy label training, we achieve state-of-the-art performance on image classification benchmarks. The code is made public at https://github.com/tmllab/PES.
****
Distributionally Robust Imitation Learning
Mohammad Ali Bashiri, Brian Ziebart, Xinhua Zhang
We consider the imitation learning problem of learning a policy in a Markov Decision Process (MDP) setting where the reward function is not given, but demonstrations from experts are available. Although the goal of imitation learning is to learn a policy that produces behaviors nearly as good as the experts’ for a desired task, assumptions of consistent optimality for demonstrated behaviors are often violated in practice. Finding a policy that is distributionally robust against noisy demonstrations based on an adversarial construction potentially solves this problem by avoiding optimistic generalizations of the demonstrated data. This paper studies Distributionally Robust Imitation Learning (DRoIL) and establishes a close connection between DRoIL and Maximum Entropy Inverse Reinforcement Learning. We show that DRoIL can be seen as a framework that maximizes a generalized concept of entropy. We develop a novel approach to transform the objective function into a convex optimization problem over a polynomial number of variables for a class of loss functions that are additive over state and action spaces. Our approach lets us optimize both stationary and non-stationary policies and, unlike prevalent previous methods, it does not require repeatedly solving an inner reinforcement learning problem. We experimentally show the significant benefits of DRoIL’s new optimization method on synthetic data and a highway driving environment.
****
On the Power of Edge Independent Graph Models
Sudhanshu Chanpuriya, Cameron Musco, Konstantinos Sotiropoulos, Charalampos Tsourakakis
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Stochastic Online Linear Regression: the Forward Algorithm to Replace Ridge
Reda Ouhamma, Odalric-Ambrym Maillard, 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.
****

Dr Jekyll & Mr Hyde: the strange case of off-policy policy updates

Romain Laroche, Remi Tachet des Combes

The policy gradient theorem states that the policy should only be updated in states that are visited by the current policy, which leads to insufficient planning in the off-policy states, and thus to convergence to suboptimal policies. We tackle this planning issue by extending the policy gradient theory to policy updates with respect to any state density. Under these generalized policy updates, we show convergence to optimality under a necessary and sufficient condition on the updates' state densities, and thereby solve the aforementioned planning issue. We also prove asymptotic convergence rates that significantly improve those in the policy gradient literature. To implement the principles prescribed by our theory, we propose an agent, Dr Jekyll & Mr Hyde (J&H), with a double personality: Dr Jekyll purely exploits while Mr Hyde purely explores. J&H's independent policies allow to record two separate replay buffers: one on-policy (Dr Jekyll's) and one off-policy (Mr Hyde's), and therefore to update J&H's models with a mixture of on-policy and off-policy updates. More than an algorithm, J&H defines principles for actor-critic algorithms to satisfy the requirements we identify in our analysis. We extensively test on finite MDPs where J&H demonstrates a superior ability to recover from converging to a suboptimal policy without impairing its speed of convergence. We also implement a deep version of the algorithm and test it on a simple problem where it shows promising results.
****

Understanding Adaptive, Multiscale Temporal Integration In Deep Speech Recognition Systems

Menoua Keshishian, Samuel Norman-Haignere, Nima Mesgarani

Natural signals such as speech are hierarchically structured across many different timescales, spanning tens (e.g., phonemes) to hundreds (e.g., words) of milliseconds, each of which is highly variable and context-dependent. While deep neural networks (DNNs) excel at recognizing complex patterns from natural signals, relatively little is known about how DNNs flexibly integrate across multiple timescales. Here, we show how a recently developed method for studying temporal integration in biological neural systems – the temporal context invariance (TCI) paradigm – can be used to understand temporal integration in DNNs. The method is simple: we measure responses to a large number of stimulus segments presented in two different contexts and estimate the smallest segment duration needed to achieve a context invariant response. We applied our method to understand how the popular DeepSpeech2 model learns to integrate across time in speech. We find that nearly all of the model units, even in recurrent layers, have a compact integration window within which stimuli substantially alter the response and outside of which stimuli have little effect. We show that training causes these integration windows to shrink at early layers and expand at higher layers, creating a hierarchy of integration windows across the network. Moreover, by measuring integration windows for time-stretched/compressed speech, we reveal a transition point, midway through the trained network, where integration windows become yoked to the duration of stimulus structures (e.g., phonemes or words) rather than absolute time. Similar phenomena were observed in a purely recurrent and purely convolutional network although structure-yoked integration was more prominent in the recurrent network. These findings suggest that deep speech recognition systems use a common motif to encode the hierarchical structure of speech: integrating across short, time-yoked windows at early layers and long, structure-yoked windows at later layers. Our method provides a straightforward and general-purpose toolkit for understanding temporal integration in black-box machine learning models.
****

VidLanKD: Improving Language Understanding via Video-Distilled Knowledge Transfer

Zineng Tang, Jaemin Cho, Hao Tan, Mohit Bansal

Since visual perception can give rich information beyond text descriptions for w
orld understanding, there has been increasing interest in leveraging visual grou
nding for language learning. Recently, vokenization (Tan and Bansal, 2020) has a
ttracted attention by using the predictions of a text-to-image retrieval model a
s labels for language model supervision. Despite its success, the method suffers
 from approximation error of using finite image labels and the lack of vocabular
y diversity of a small image-text dataset. To overcome these limitations, we pre
sent VidLanKD, a video-language knowledge distillation method for improving lang
uage understanding. We train a multi-modal teacher model on a video-text dataset
, and then transfer its knowledge to a student language model with a text datase
t. To avoid approximation error, we propose to use different knowledge distillat
ion objectives. In addition, the use of a large-scale video-text dataset helps l
earn diverse and richer vocabularies. In our experiments, VidLanKD achieves cons
istent improvements over text-only language models and vokenization models, on s
everal downstream language understanding tasks including GLUE, SQuAD, and SWAG.
We also demonstrate the improved world knowledge, physical reasoning, and tempor
al reasoning capabilities of our model by evaluating on the GLUE-diagnostics, PI
QA, and TRACIE datasets. Lastly, we present comprehensive ablation studies as we
ll as visualizations of the learned text-to-video grounding results of our teach
er and student language models.
****
Detecting Individual Decision-Making Style: Exploring Behavioral Stylometry in C
hess
Reid McIlroy-Young, Yu Wang, Siddhartha Sen, Jon Kleinberg, Ashton Anderson
The advent of machine learning models that surpass human decision-making ability
 in complex domains has initiated a movement towards building AI systems that in
teract with humans. Many building blocks are essential for this activity, with a
 central one being the algorithmic characterization of human behavior. While muc
h of the existing work focuses on aggregate human behavior, an important long-ra
nge goal is to develop behavioral models that specialize to individual people an
d can differentiate among them.To formalize this process, we study the problem o
f behavioral stylometry, in which the task is to identify a decision-maker from
their decisions alone. We present a transformer-based approach to behavioral sty
lometry in the context of chess, where one attempts to identify the player who p
layed a set of games. Our method operates in a few-shot classification framework
, and can correctly identify a player from among thousands of candidate players
with 98% accuracy given only 100 labeled games. Even when trained on amateur pla
y, our method generalises to out-of-distribution samples of Grandmaster players,
 despite the dramatic differences between amateur and world-class players. Final
ly, we consider more broadly what our resulting embeddings reveal about human st
yle in chess, as well as the potential ethical implications of powerful methods
for identifying individuals from behavioral data.
****
Coupled Gradient Estimators for Discrete Latent Variables
Zhe Dong, Andriy Mnih, George Tucker
Training models with discrete latent variables is challenging due to the high va
riance of unbiased gradient estimators. While low-variance reparameterization gr
adients of a continuous relaxation can provide an effective solution, a continuo
us relaxation is not always available or tractable. Dong et al. (2020) and Yin e
t al. (2020) introduced a performant estimator that does not rely on continuous
relaxations; however, it is limited to binary random variables. We introduce a n
ovel derivation of their estimator based on importance sampling and statistical
couplings, which we extend to the categorical setting. Motivated by the construc
tion of a stick-breaking coupling, we introduce gradient estimators based on rep
arameterizing categorical variables as sequences of binary variables and Rao-Bla
ckwellization. In systematic experiments, we show that our proposed categorical
gradient estimators provide state-of-the-art performance, whereas even with addi
tional Rao-Blackwellization previous estimators (Yin et al., 2019) underperform
a simpler REINFORCE with a leave-one-out-baseline estimator (Kool et al., 2019).
****

AutoGEL: An Automated Graph Neural Network with Explicit Link Information
Zhili Wang, Shimin DI, Lei Chen
Recently, Graph Neural Networks (GNNs) have gained popularity in a variety of real-world scenarios. Despite the great success, the architecture design of GNNs heavily relies on manual labor. Thus, automated graph neural network (AutoGNN) has attracted interest and attention from the research community, which makes significant performance improvements in recent years. However, existing AutoGNN works mainly adopt an implicit way to model and leverage the link information in the graphs, which is not well regularized to the link prediction task on graphs, and limits the performance of AutoGNN for other graph tasks. In this paper, we present a novel AutoGNN work that explicitly models the link information, abbreviated to AutoGEL. In such a way, AutoGEL can handle the link prediction task and improve the performance of AutoGNNs on the node classification and graph classification task. Moreover, AutoGEL proposes a novel search space containing various design dimensions at both intra-layer and inter-layer designs and adopts a more robust differentiable search algorithm to further improve efficiency and effectiveness. Experimental results on benchmark data sets demonstrate the superiority of AutoGEL on several tasks.
****
RL for Latent MDPs: Regret Guarantees and a Lower Bound
Jeongyeol Kwon, Yonathan Efroni, Constantine Caramanis, 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-authors prior to requesting a name change in the electronic proceedings.
****
Adaptive Sampling for Minimax Fair Classification
Shubhanshu Shekhar, Greg Fields, Mohammad Ghavamzadeh, Tara Javidi
Machine learning models trained on uncurated datasets can often end up adversely affecting inputs belonging to underrepresented groups. To address this issue, we consider the problem of adaptively constructing training sets which allow us to learn classifiers that are fair in a {\em minimax} sense. We first propose an adaptive sampling algorithm based on the principle of \emph{optimism}, and derive theoretical bounds on its performance. We also propose heuristic extensions of this algorithm suitable for application to large scale, practical problems. Next, by deriving algorithm independent lower-bounds for a specific class of problems, we show that the performance achieved by our adaptive scheme cannot be improved in general. We then validate the benefits of adaptively constructing training sets via experiments on synthetic tasks with logistic regression classifiers, as well as on several real-world tasks using convolutional neural networks (CNNs).
****
Structured in Space, Randomized in Time: Leveraging Dropout in RNNs for Efficient Training
Anup Sarma, Sonali Singh, Huaipan Jiang, Rui Zhang, Mahmut Kandemir, Chita Das
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Variational Continual Bayesian Meta-Learning
Qiang Zhang, Jinyuan Fang, Zaiqiao Meng, Shangsong Liang, Emine Yilmaz
Conventional meta-learning considers a set of tasks from a stationary distribution. In contrast, this paper focuses on a more complex online setting, where tasks arrive sequentially and follow a non-stationary distribution. Accordingly, we propose a Variational Continual Bayesian Meta-Learning (VC-BML) algorithm. VC-BML maintains a Dynamic Gaussian Mixture Model for meta-parameters, with the number of component distributions determined by a Chinese Restaurant Process. Dynamic mixtures at the meta-parameter level increase the capability to adapt to diverse tasks due to a larger parameter space, alleviating the negative knowledge tran

sfer problem. To infer posteriors of model parameters, compared to the previously used point estimation method, we develop a more robust posterior approximation method -- structured variational inference for the sake of avoiding forgetting knowledge. Experiments on tasks from non-stationary distributions show that VC-BML is superior in transferring knowledge among diverse tasks and alleviating catastrophic forgetting in an online setting.
****
Recognizing Vector Graphics without Rasterization
XINYANG JIANG, LU LIU, Caihua Shan, Yifei Shen, Xuanyi Dong, Dongsheng Li
In this paper, we consider a different data format for images: vector graphics. In contrast to raster graphics which are widely used in image recognition, vector graphics can be scaled up or down into any resolution without aliasing or information loss, due to the analytic representation of the primitives in the document. Furthermore, vector graphics are able to give extra structural information on how low-level elements group together to form high level shapes or structures. These merits of graphic vectors have not been fully leveraged in existing methods. To explore this data format, we target on the fundamental recognition tasks: object localization and classification. We propose an efficient CNN-free pipeline that does not render the graphic into pixels (i.e. rasterization), and takes textual document of the vector graphics as input, called YOLaT (You Only Look at Text). YOLaT builds multi-graphs to model the structural and spatial information in vector graphics, and a dual-stream graph neural network is proposed to detect objects from the graph. Our experiments show that by directly operating on vector graphics, YOLaT outperforms raster-graphic based object detection baselines in terms of both average precision and efficiency. Code is available at https://github.com/microsoft/YOLaT-VectorGraphicsRecognition.
****
On Episodes, Prototypical Networks, and Few-Shot Learning
Steinar Laenen, Luca Bertinetto
Episodic learning is a popular practice among researchers and practitioners interested in few-shot learning.It consists of organising training in a series of learning problems (or episodes), each divided into a small training and validation subset to mimic the circumstances encountered during evaluation.But is this always necessary?In this paper, we investigate the usefulness of episodic learning in methods which use nonparametric approaches, such as nearest neighbours, at the level of the episode.For these methods, we not only show how the constraints imposed by episodic learning are not necessary, but that they in fact lead to a data-inefficient way of exploiting training batches.We conduct a wide range of ablative experiments with Matching and Prototypical Networks, two of the most popular methods that use nonparametric approaches at the level of the episode.Their "non-episodic'' counterparts are considerably simpler, have less hyperparameters, and improve their performance in multiple few-shot classification datasets.
****
Pointwise Bounds for Distribution Estimation under Communication Constraints
Wei-Ning Chen, Peter Kairouz, Ayfer Ozgur
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
CHIP: CHannel Independence-based Pruning for Compact Neural Networks
Yang Sui, Miao Yin, Yi Xie, Huy Phan, Saman Aliari Zonouz, Bo Yuan
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Federated Split Task-Agnostic Vision Transformer for COVID-19 CXR Diagnosis
Sangjoon Park, Gwanghyun Kim, Jeongsol Kim, Boah Kim, Jong Chul Ye
Federated learning, which shares the weights of the neural network across client

s, is gaining attention in the healthcare sector as it enables training on a lar ge corpus of decentralized data while maintaining data privacy. For example, thi s enables neural network training for COVID-19 diagnosis on chest X-ray (CXR) im ages without collecting patient CXR data across multiple hospitals. Unfortunatel y, the exchange of the weights quickly consumes the network bandwidth if highly expressive network architecture is employed. So-called split learning partially solves this problem by dividing a neural network into a client and a server part , so that the client part of the network takes up less extensive computation res ources and bandwidth. However, it is not clear how to find the optimal split wit hout sacrificing the overall network performance. To amalgamate these methods an d thereby maximize their distinct strengths, here we show that the Vision Transf ormer, a recently developed deep learning architecture with straightforward deco mposable configuration, is ideally suitable for split learning without sacrifici ng performance. Even under the non-independent and identically distributed data distribution which emulates a real collaboration between hospitals using CXR dat asets from multiple sources, the proposed framework was able to attain performan ce comparable to data-centralized training. In addition, the proposed framework along with heterogeneous multi-task clients also improves individual task perfor mances including the diagnosis of COVID-19, eliminating the need for sharing lar ge weights with innumerable parameters. Our results affirm the suitability of Tr ansformer for collaborative learning in medical imaging and pave the way forward for future real-world implementations.

****

Active Offline Policy Selection

Ksenia Konyushova, Yutian Chen, Thomas Paine, Caglar Gulcehre, Cosmin Paduraru, Daniel J. Mankowitz, Misha Denil, Nando de Freitas

This paper addresses the problem of policy selection in domains with abundant lo gged data, but with a restricted interaction budget. Solving this problem would enable safe evaluation and deployment of offline reinforcement learning policies in industry, robotics, and recommendation domains among others. Several off-pol icy evaluation (OPE) techniques have been proposed to assess the value of polici es using only logged data. However, there is still a big gap between the evaluat ion by OPE and the full online evaluation in the real environment. Yet, large am ounts of online interactions are often not possible in practice. To overcome thi s problem, we introduce active offline policy selection --- a novel sequential d ecision approach that combines logged data with online interaction to identify t he best policy. This approach uses OPE estimates to warm start the online evalua tion. Then, in order to utilize the limited environment interactions wisely we d ecide which policy to evaluate next based on a Bayesian optimization method with a kernel function that represents policy similarity. We use multiple benchmarks with a large number of candidate policies to show that the proposed approach im proves upon state-of-the-art OPE estimates and pure online policy evaluation.

****

Unsupervised Representation Transfer for Small Networks: I Believe I Can Distill On-the-Fly

Hee Min Choi, Hyoa Kang, Dokwan Oh

A current remarkable improvement of unsupervised visual representation learning is based on heavy networks with large-batch training. While recent methods have greatly reduced the gap between supervised and unsupervised performance of deep models such as ResNet-50, this development has been relatively limited for small models. In this work, we propose a novel unsupervised learning framework for sm all networks that combines deep self-supervised representation learning and know ledge distillation within one-phase training. In particular, a teacher model is trained to produce consistent cluster assignments between different views of the same image. Simultaneously, a student model is encouraged to mimic the predicti on of on-the-fly self-supervised teacher. For effective knowledge transfer, we a dopt the idea of domain classifier so that student training is guided by discrim inative features invariant to the representational space shift between teacher a nd student. We also introduce a network driven multi-view generation paradigm to capture rich feature information contained in the network itself. Extensive exp

eriments show that our student models surpass state-of-the-art offline distilled networks even from stronger self-supervised teachers as well as top-performing self-supervised models. Notably, our ResNet-18, trained with ResNet-50 teacher, achieves 68.3% ImageNet Top-1 accuracy on frozen feature linear evaluation, which is only 1.5% below the supervised baseline.
****

Understanding Bandits with Graph Feedback

Houshuang Chen, zengfeng Huang, Shuai Li, Chihao 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.
****

Information-theoretic generalization bounds for black-box learning algorithms

Hrayr Harutyunyan, Maxim Raginsky, Greg Ver Steeg, Aram Galstyan

We derive information-theoretic generalization bounds for supervised learning algorithms based on the information contained in predictions rather than in the output of the training algorithm. These bounds improve over the existing information-theoretic bounds, are applicable to a wider range of algorithms, and solve two key challenges: (a) they give meaningful results for deterministic algorithms and (b) they are significantly easier to estimate. We show experimentally that the proposed bounds closely follow the generalization gap in practical scenarios for deep learning.
****

Trash or Treasure? An Interactive Dual-Stream Strategy for Single Image Reflection Separation

Qiming Hu, Xiaojie Guo

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

Rot-Pro: Modeling Transitivity by Projection in Knowledge Graph Embedding

Tengwei Song, Jie Luo, Lei Huang

Knowledge graph embedding models learn the representations of entities and relations in the knowledge graphs for predicting missing links (relations) between entities. Their effectiveness are deeply affected by the ability of modeling and inferring different relation patterns such as symmetry, asymmetry, inversion, composition and transitivity. Although existing models are already able to model many of these relations patterns, transitivity, a very common relation pattern, is still not been fully supported. In this paper, we first theoretically show that the transitive relations can be modeled with projections. We then propose the Rot-Pro model which combines the projection and relational rotation together. We prove that Rot-Pro can infer all the above relation patterns. Experimental results show that the proposed Rot-Pro model effectively learns the transitivity pattern and achieves the state-of-the-art results on the link prediction task in the datasets containing transitive relations.
****

Planning from Pixels in Environments with Combinatorially Hard Search Spaces

Marco Bagatella, Miroslav Olšák, Michal Rolínek, Georg Martius

The ability to form complex plans based on raw visual input is a litmus test for current capabilities of artificial intelligence, as it requires a seamless combination of visual processing and abstract algorithmic execution, two traditionally separate areas of computer science. A recent surge of interest in this field brought advances that yield good performance in tasks ranging from arcade games to continuous control; these methods however do not come without significant issues, such as limited generalization capabilities and difficulties when dealing with combinatorially hard planning instances. Our contribution is two-fold: (i) we present a method that learns to represent its environment as a latent graph and leverages state reidentification to reduce the complexity of finding a good po

licy from exponential to linear (ii) we introduce a set of lightweight environme
nts with an underlying discrete combinatorial structure in which planning is cha
llenging even for humans. Moreover, we show that our methods achieves strong emp
irical generalization to variations in the environment, even across highly disad
vantaged regimes, such as “one-shot” planning, or in an offline RL paradigm whic
h only provides low-quality trajectories.
****
PLUGIn: A simple algorithm for inverting generative models with recovery guarant
ees
Babhru Joshi, Xiaowei Li, Yaniv Plan, Ozgur Yilmaz
Requests for name changes in the electronic proceedings will be accepted with 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.
****
Modular Gaussian Processes for Transfer Learning
Pablo Moreno-Muñoz, Antonio Artes, Mauricio Álvarez
We present a framework for transfer learning based on modular variational Gaussi
an processes (GP). We develop a module-based method that having a dictionary of
well fitted GPs, each model being characterised by its hyperparameters, pseudo-i
nputs and their corresponding posterior densities, one could build ensemble GP m
odels without revisiting any data. Our method avoids undesired data centralisati
on, reduces rising computational costs and allows the transfer of learned uncert
ainty metrics after training. We exploit the augmentation of high-dimensional in
tegral operators based on the Kullback-Leibler divergence between stochastic pro
cesses to introduce an efficient lower bound under all the sparse variational GP
s, with different complexity and even likelihood distribution. The method is als
o valid for multi-output GPs, learning correlations a posteriori between indepen
dent modules. Extensive results illustrate the usability of our framework in lar
ge-scale and multi-task experiments, also compared with the exact inference meth
ods in the literature.
****
Neural Human Performer: Learning Generalizable Radiance Fields for Human Perform
ance Rendering
Youngjoong Kwon, Dahun Kim, Duygu Ceylan, Henry Fuchs
In this paper, we aim at synthesizing a free-viewpoint video of an arbitrary hum
an performance using sparse multi-view cameras. Recently, several works have add
ressed this problem by learning person-specific neural radiance fields (NeRF) to
 capture the appearance of a particular human. In parallel, some work proposed t
o use pixel-aligned features to generalize radiance fields to arbitrary new scen
es and objects. Adopting such generalization approaches to humans, however, is h
ighly challenging due to the heavy occlusions and dynamic articulations of body
parts. To tackle this, we propose Neural Human Performer, a novel approach that
learns generalizable neural radiance fields based on a parametric human body mod
el for robust performance capture. Specifically, we first introduce a temporal t
ransformer that aggregates tracked visual features based on the skeletal body mo
tion over time. Moreover, a multi-view transformer is proposed to perform cross-
attention between the temporally-fused features and the pixel-aligned features a
t each time step to integrate observations on the fly from multiple views. Exper
iments on the ZJU-MoCap and AIST datasets show that our method significantly out
performs recent generalizable NeRF methods on unseen identities and poses.
****
Locally differentially private estimation of functionals of discrete distributio
ns
Cristina Butucea, Yann ISSARTEL
Requests for name changes in the electronic proceedings will be accepted with 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.
****

Asymptotics of representation learning in finite Bayesian neural networks
Jacob Zavatone-Veth, Abdulkadir Canatar, Ben Ruben, Cengiz Pehlevan
Recent works have suggested that finite Bayesian neural networks may sometimes outperform their infinite cousins because finite networks can flexibly adapt their internal representations. However, our theoretical understanding of how the learned hidden layer representations of finite networks differ from the fixed representations of infinite networks remains incomplete. Perturbative finite-width corrections to the network prior and posterior have been studied, but the asymptotics of learned features have not been fully characterized. Here, we argue that the leading finite-width corrections to the average feature kernels for any Bayesian network with linear readout and Gaussian likelihood have a largely universal form. We illustrate this explicitly for three tractable network architectures: deep linear fully-connected and convolutional networks, and networks with a single nonlinear hidden layer. Our results begin to elucidate how task-relevant learning signals shape the hidden layer representations of wide Bayesian neural networks.
****
Adaptive Ensemble Q-learning: Minimizing Estimation Bias via Error Feedback
Hang Wang, Sen Lin, Junshan Zhang
The ensemble method is a promising way to mitigate the overestimation issue in Q-learning, where multiple function approximators are used to estimate the action values. It is known that the estimation bias hinges heavily on the ensemble size (i.e., the number of Q-function approximators used in the target), and that determining the 'right' ensemble size is highly nontrivial, because of the time-varying nature of the function approximation errors during the learning process. To tackle this challenge, we first derive an upper bound and a lower bound on the estimation bias, based on which the ensemble size is adapted to drive the bias to be nearly zero, thereby coping with the impact of the time-varying approximation errors accordingly. Motivated by the theoretic findings, we advocate that the ensemble method can be combined with Model Identification Adaptive Control (MIAC) for effective ensemble size adaptation. Specifically, we devise Adaptive Ensemble Q-learning (AdaEQ), a generalized ensemble method with two key steps: (a) approximation error characterization which serves as the feedback for flexibly controlling the ensemble size, and (b) ensemble size adaptation tailored towards minimizing the estimation bias.   Extensive experiments are carried out to show that AdaEQ can improve the learning performance than the existing methods for the MuJoCo benchmark.
****
Domain Adaptation with Invariant Representation Learning: What Transformations to Learn?
Petar Stojanov, Zijian Li, Mingming Gong, Ruichu Cai, Jaime Carbonell, 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-authors prior to requesting a name change in the electronic proceedings.
****
CSDI: Conditional Score-based Diffusion Models for Probabilistic Time Series Imputation
Yusuke Tashiro, Jiaming Song, Yang Song, Stefano Ermon
The imputation of missing values in time series has many applications in healthcare and finance. While autoregressive models are natural candidates for time series imputation, score-based diffusion models have recently outperformed existing counterparts including autoregressive models in many tasks such as image generation and audio synthesis, and would be promising for time series imputation. In this paper, we propose Conditional Score-based Diffusion model (CSDI), a novel time series imputation method that utilizes score-based diffusion models conditioned on observed data. Unlike existing score-based approaches, the conditional diffusion model is explicitly trained for imputation and can exploit correlations between observed values. On healthcare and environmental data, CSDI improves by 40-65% over existing probabilistic imputation methods on popular performance met

rics. In addition, deterministic imputation by CSDI reduces the error by 5-20% c ompared to the state-of-the-art deterministic imputation methods. Furthermore, C SDI can also be applied to time series interpolation and probabilistic forecasti ng, and is competitive with existing baselines. The code is available at https:/ /github.com/ermongroup/CSDI.
****
Causal Bandits with Unknown Graph Structure
Yangyi Lu, Amirhossein Meisami, Ambuj Tewari
In causal bandit problems the action set consists of interventions on variables of a causal graph. Several researchers have recently studied such bandit problem s and pointed out their practical applications. However, all existing works rely on a restrictive and impractical assumption that the learner is given full know ledge of the causal graph structure upfront. In this paper, we develop novel cau sal bandit algorithms without knowing the causal graph. Our algorithms work well for causal trees, causal forests and a general class of causal graphs. The regr et guarantees of our algorithms greatly improve upon those of standard multi-ar med bandit (MAB) algorithms under mild conditions. Lastly, we prove our mild con ditions are necessary: without them one cannot do better than standard MAB algor ithms.
****
Piper: Multidimensional Planner for DNN Parallelization
Jakub M. Tarnawski, Deepak Narayanan, Amar Phanishayee
The rapid increase in sizes of state-of-the-art DNN models, and consequently the increase in the compute and memory requirements of model training, has led to t he development of many execution schemes such as data parallelism, pipeline mode l parallelism, tensor (intra-layer) model parallelism, and various memory-saving optimizations. However, no prior work has tackled the highly complex problem of optimally partitioning the DNN computation graph across many accelerators while combining all these parallelism modes and optimizations.In this work, we introd uce Piper, an efficient optimization algorithm for this problem that is based on a two-level dynamic programming approach. Our two-level approach is driven by t he insight that being given tensor-parallelization techniques for individual lay ers (e.g., Megatron-LM's splits for transformer layers) significantly reduces th e search space and makes the global problem tractable, compared to considering t ensor-parallel configurations for the entire DNN operator graph.
****
Causal Effect Inference for Structured Treatments
Jean Kaddour, Yuchen Zhu, Qi Liu, Matt J. Kusner, Ricardo Silva
We address the estimation of conditional average treatment effects (CATEs) for s tructured treatments (e.g., graphs, images, texts). Given a weak condition on th e effect, we propose the generalized Robinson decomposition, which (i) isolates the causal estimand (reducing regularization bias), (ii) allows one to plug in a rbitrary models for learning, and (iii) possesses a quasi-oracle convergence gua rantee under mild assumptions. In experiments with small-world and molecular gra phs we demonstrate that our approach outperforms prior work in CATE estimation.
****
Efficient hierarchical Bayesian inference for spatio-temporal regression models in neuroimaging
Ali Hashemi, Yijing Gao, Chang Cai, Sanjay Ghosh, Klaus-Robert Müller, Srikantan Nagarajan, Stefan Haufe
Several problems in neuroimaging and beyond require inference on the parameters of multi-task sparse hierarchical regression models. Examples include M/EEG inve rse problems, neural encoding models for task-based fMRI analyses, and climate s cience. In these domains, both the model parameters to be inferred and the measu rement noise may exhibit a complex spatio-temporal structure. Existing work eith er neglects the temporal structure or leads to computationally demanding inferen ce schemes. Overcoming these limitations, we devise a novel flexible hierarchica l Bayesian framework within which the spatio-temporal dynamics of model paramete rs and noise are modeled to have Kronecker product covariance structure. Inferen ce in our framework is based on majorization-minimization optimization and has g

uaranteed convergence properties. Our highly efficient algorithms exploit the intrinsic Riemannian geometry of temporal autocovariance matrices. For stationary dynamics described by Toeplitz matrices, the theory of circulant embeddings is employed. We prove convex bounding properties and derive update rules of the resulting algorithms. On both synthetic and real neural data from M/EEG, we demonstrate that our methods lead to improved performance.
****

Topological Attention for Time Series Forecasting
Sebastian Zeng, Florian Graf, Christoph Hofer, Roland Kwitt
The problem of (point) forecasting univariate time series is considered. Most approaches, ranging from traditional statistical methods to recent learning-based techniques with neural networks, directly operate on raw time series observations. As an extension, we study whether local topological properties, as captured via persistent homology, can serve as a reliable signal that provides complementary information for learning to forecast. To this end, we propose topological attention, which allows attending to local topological features within a time horizon of historical data. Our approach easily integrates into existing end-to-end trainable forecasting models, such as N-BEATS, and, in combination with the latter exhibits state-of-the-art performance on the large-scale M4 benchmark dataset of 100,000 diverse time series from different domains. Ablation experiments, as well as a comparison to recent techniques in a setting where only a single time series is available for training, corroborate the beneficial nature of including local topological information through an attention mechanism.
****

Local Signal Adaptivity: Provable Feature Learning in Neural Networks Beyond Kernels
Stefani Karp, Ezra Winston, Yuanzhi Li, Aarti Singh
Neural networks have been shown to outperform kernel methods in practice (including neural tangent kernels). Most theoretical explanations of this performance gap focus on learning a complex hypothesis class; in some cases, it is unclear whether this hypothesis class captures realistic data. In this work, we propose a related, but alternative, explanation for this performance gap in the image classification setting, based on finding a sparse signal in the presence of noise. Specifically, we prove that, for a simple data distribution with sparse signal amidst high-variance noise, a simple convolutional neural network trained using stochastic gradient descent learns to threshold out the noise and find the signal. On the other hand, the corresponding neural tangent kernel, with a fixed set of predetermined features, is unable to adapt to the signal in this manner. We supplement our theoretical results by demonstrating this phenomenon empirically: in CIFAR-10 and MNIST images with various backgrounds, as the background noise increases in intensity, a CNN's performance stays relatively robust, whereas its corresponding neural tangent kernel sees a notable drop in performance. We therefore propose the "local signal adaptivity" (LSA) phenomenon as one explanation for the superiority of neural networks over kernel methods.
****

IA-RED^2: Interpretability-Aware Redundancy Reduction for Vision Transformers
Bowen Pan, Rameswar Panda, Yifan Jiang, Zhangyang Wang, Rogerio Feris, Aude Oliva
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

Symbolic Regression via Deep Reinforcement Learning Enhanced Genetic Programming Seeding
Terrell Mundhenk, Mikel Landajuela, Ruben Glatt, Claudio P Santiago, Daniel faissol, Brenden K Petersen
Symbolic regression is the process of identifying mathematical expressions that fit observed output from a black-box process. It is a discrete optimization problem generally believed to be NP-hard. Prior approaches to solving the problem in

clude neural-guided search (e.g. using reinforcement learning) and genetic programming. In this work, we introduce a hybrid neural-guided/genetic programming approach to symbolic regression and other combinatorial optimization problems. We propose a neural-guided component used to seed the starting population of a random restart genetic programming component, gradually learning better starting populations. On a number of common benchmark tasks to recover underlying expressions from a dataset, our method recovers 65% more expressions than a recently published top-performing model using the same experimental setup. We demonstrate that running many genetic programming generations without interdependence on the neural-guided component performs better for symbolic regression than alternative formulations where the two are more strongly coupled. Finally, we introduce a new set of 22 symbolic regression benchmark problems with increased difficulty over existing benchmarks. Source code is provided at www.github.com/brendenpetersen/deep-symbolic-optimization.
****

Choose a Transformer: Fourier or Galerkin
Shuhao Cao
In this paper, we apply the self-attention from the state-of-the-art Transformer in Attention Is All You Need for the first time to a data-driven operator learning problem related to partial differential equations. An effort is put together to explain the heuristics of, and to improve the efficacy of the attention mechanism. By employing the operator approximation theory in Hilbert spaces, it is demonstrated for the first time that the softmax normalization in the scaled dot-product attention is sufficient but not necessary. Without softmax, the approximation capacity of a linearized Transformer variant can be proved to be comparable to a Petrov-Galerkin projection layer-wise, and the estimate is independent with respect to the sequence length. A new layer normalization scheme mimicking the Petrov-Galerkin projection is proposed to allow a scaling to propagate through attention layers, which helps the model achieve remarkable accuracy in operator learning tasks with unnormalized data. Finally, we present three operator learning experiments, including the viscid Burgers' equation, an interface Darcy flow, and an inverse interface coefficient identification problem. The newly proposed simple attention-based operator learner, Galerkin Transformer, shows significant improvements in both training cost and evaluation accuracy over its softmax-normalized counterparts.
****

A Causal Lens for Controllable Text Generation
Zhiting Hu, Li Erran Li
Controllable text generation concerns two fundamental tasks of wide applications, namely generating text of given attributes (i.e., attribute-conditional generation), and minimally editing existing text to possess desired attributes (i.e., text attribute transfer). Extensive prior work has largely studied the two problems separately, and developed different conditional models which, however, are prone to producing biased text (e.g., various gender stereotypes). This paper proposes to formulate controllable text generation from a principled causal perspective which models the two tasks with a unified framework. A direct advantage of the causal formulation is the use of rich causality tools to mitigate generation biases and improve control. We treat the two tasks as interventional and counterfactual causal inference based on a structural causal model, respectively. We then apply the framework to the challenging practical setting where confounding factors (that induce spurious correlations) are observable only on a small fraction of data. Experiments show significant superiority of the causal approach over previous conditional models for improved control accuracy and reduced bias.
****

Differentially Private Multi-Armed Bandits in the Shuffle Model
Jay Tenenbaum, Haim Kaplan, Yishay Mansour, Uri Stemmer
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 Adaptivity: A Universal Algorithm for Minimizing the Adaptive Regret of Con vex Functions
Lijun Zhang, Guanghui Wang, Wei-Wei Tu, Wei Jiang, Zhi-Hua Zhou
To deal with changing environments, a new performance measure—adaptive regret, d efined as the maximum static regret over any interval, was proposed in online le arning. Under the setting of online convex optimization, several algorithms have been successfully developed to minimize the adaptive regret. However, existing algorithms lack universality in the sense that they can only handle one type of convex functions and need apriori knowledge of parameters. By contrast, there ex ist universal algorithms, such as MetaGrad, that attain optimal static regret fo r multiple types of convex functions simultaneously. Along this line of research , this paper presents the first universal algorithm for minimizing the adaptive regret of convex functions. Specifically, we borrow the idea of maintaining mult iple learning rates in MetaGrad to handle the uncertainty of functions, and util ize the technique of sleeping experts to capture changing environments. In this way, our algorithm automatically adapts to the property of functions (convex, ex ponentially concave, or strongly convex), as well as the nature of environments (stationary or changing). As a by product, it also allows the type of functions to switch between rounds.
****
Learning Hard Optimization Problems: A Data Generation Perspective
James Kotary, Ferdinando Fioretto, Pascal Van Hentenryck
Optimization problems are ubiquitous in our societies and are present in almost every segment of the economy. Most of these optimization problems are NP-hard an d computationally demanding, often requiring approximate solutions for large-sca le instances. Machine learning frameworks that learn to approximate solutions to such hard optimization problems are a potentially promising avenue to address t hese difficulties, particularly when many closely related problem instances must be solved repeatedly. Supervised learning frameworks can train a model using th e outputs of pre-solved instances. However, when the outputs are themselves appr oximations, when the optimization problem has symmetric solutions, and/or when t he solver uses randomization, solutions to closely related instances may exhibit large differences and the learning task can become inherently more difficult. T his paper demonstrates this critical challenge, connects the volatility of the t raining data to the ability of a model to approximate it, and proposes a method for producing (exact or approximate) solutions to optimization problems that are more amenable to supervised learning tasks. The effectiveness of the method is tested on hard non-linear nonconvex and discrete combinatorial problems.
****
Canonical Capsules: Self-Supervised Capsules in Canonical Pose
Weiwei Sun, Andrea Tagliasacchi, Boyang Deng, Sara Sabour, Soroosh Yazdani, Geof frey E. Hinton, Kwang Moo Yi
We propose a self-supervised capsule architecture for 3D point clouds. We comput e capsule decompositions of objects through permutation-equivariant attention, a nd self-supervise the process by training with pairs of randomly rotated objects . Our key idea is to aggregate the attention masks into semantic keypoints, and use these to supervise a decomposition that satisfies the capsule invariance/equ ivariance properties. This not only enables the training of a semantically consi stent decomposition, but also allows us to learn a canonicalization operation th at enables object-centric reasoning. To train our neural network we require neit her classification labels nor manually-aligned training datasets. Yet, by learni ng an object-centric representation in a self-supervised manner, our method outp erforms the state-of-the-art on 3D point cloud reconstruction, canonicalization, and unsupervised classification.
****
Characterizing Generalization under Out-Of-Distribution Shifts in Deep Metric Le arning
Timo Milbich, Karsten Roth, Samarth Sinha, Ludwig Schmidt, Marzyeh Ghassemi, Bjo rn Ommer

Deep Metric Learning (DML) aims to find representations suitable for zero-shot transfer to a priori unknown test distributions. However, common evaluation protocols only test a single, fixed data split in which train and test classes are assigned randomly. More realistic evaluations should consider a broad spectrum of distribution shifts with potentially varying degree and difficulty.In this work, we systematically construct train-test splits of increasing difficulty and present the ooDML benchmark to characterize generalization under out-of-distribution shifts in DML. ooDML is designed to probe the generalization performance on much more challenging, diverse train-to-test distribution shifts. Based on our new benchmark, we conduct a thorough empirical analysis of state-of-the-art DML methods. We find that while generalization tends to consistently degrade with difficulty, some methods are better at retaining performance as the distribution shift increases. Finally, we propose few-shot DML as an efficient way to consistently improve generalization in response to unknown test shifts presented in ooDML.
****

Dynamics-regulated kinematic policy for egocentric pose estimation
Zhengyi Luo, Ryo Hachiuma, Ye Yuan, Kris Kitani
We propose a method for object-aware 3D egocentric pose estimation that tightly integrates kinematics modeling, dynamics modeling, and scene object information. Unlike prior kinematics or dynamics-based approaches where the two components are used disjointly, we synergize the two approaches via dynamics-regulated training. At each timestep, a kinematic model is used to provide a target pose using video evidence and simulation state. Then, a prelearned dynamics model attempts to mimic the kinematic pose in a physics simulator. By comparing the pose instructed by the kinematic model against the pose generated by the dynamics model, we can use their misalignment to further improve the kinematic model. By factoring in the 6DoF pose of objects (e.g., chairs, boxes) in the scene, we demonstrate for the first time, the ability to estimate physically-plausible 3D human-object interactions using a single wearable camera. We evaluate our egocentric pose estimation method in both controlled laboratory settings and real-world scenarios.
****

Never Go Full Batch (in Stochastic Convex Optimization)
Idan Amir, Yair Carmon, Tomer Koren, Roi Livni
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

Collaborative Learning in the Jungle (Decentralized, Byzantine, Heterogeneous, Asynchronous and Nonconvex Learning)
El Mahdi El-Mhamdi, Sadegh Farhadkhani, Rachid Guerraoui, Arsany Guirguis, Lê-Nguyên Hoang, Sébastien Rouault
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

Not All Low-Pass Filters are Robust in Graph Convolutional Networks
Heng Chang, Yu Rong, Tingyang Xu, Yatao Bian, Shiji Zhou, Xin Wang, Junzhou Huang, Wenwu Zhu
Graph Convolutional Networks (GCNs) are promising deep learning approaches in learning representations for graph-structured data. Despite the proliferation of such methods, it is well known that they are vulnerable to carefully crafted adversarial attacks on the graph structure. In this paper, we first conduct an adversarial vulnerability analysis based on matrix perturbation theory. We prove that the low- frequency components of the symmetric normalized Laplacian, which is usually used as the convolutional filter in GCNs, could be more robust against structural perturbations when their eigenvalues fall into a certain robust interval. Our results indicate that not all low-frequency components are robust to adversarial attacks and provide a deeper understanding of the relationship between g

raph spectrum and robustness of GCNs. Motivated by the theory, we present GCN-LFR, a general robust co-training paradigm for GCN-based models, that encourages transferring the robustness of low-frequency components with an auxiliary neural network. To this end, GCN-LFR could enhance the robustness of various kinds of GCN-based models against poisoning structural attacks in a plug-and-play manner. Extensive experiments across five benchmark datasets and five GCN-based models also confirm that GCN-LFR is resistant to the adversarial attacks without compromising on performance in the benign situation.
****
Counterfactual Maximum Likelihood Estimation for Training Deep Networks
Xinyi Wang, Wenhu Chen, Michael Saxon, William Yang Wang
Although deep learning models have driven state-of-the-art performance on a wide array of tasks, they are prone to spurious correlations that should not be learned as predictive clues. To mitigate this problem, we propose a causality-based training framework to reduce the spurious correlations caused by observed confounders. We give theoretical analysis on the underlying general Structural Causal Model (SCM) and propose to perform Maximum Likelihood Estimation (MLE) on the interventional distribution instead of the observational distribution, namely Counterfactual Maximum Likelihood Estimation (CMLE). As the interventional distribution, in general, is hidden from the observational data, we then derive two different upper bounds of the expected negative log-likelihood and propose two general algorithms, Implicit CMLE and Explicit CMLE, for causal predictions of deep learning models using observational data. We conduct experiments on both simulated data and two real-world tasks: Natural Language Inference (NLI) and Image Captioning. The results show that CMLE methods outperform the regular MLE method in terms of out-of-domain generalization performance and reducing spurious correlations, while maintaining comparable performance on the regular evaluations.
****
Robust Optimization for Multilingual Translation with Imbalanced Data
Xian Li, Hongyu Gong
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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/B/n Testing with Control in the Presence of Subpopulations
Yoan Russac, Christina Katsimerou, Dennis Bohle, Olivier Cappé, Aurélien Garivier, Wouter M. Koolen
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Using Random Effects to Account for High-Cardinality Categorical Features and Repeated Measures in Deep Neural Networks
Giora Simchoni, Saharon Rosset
High-cardinality categorical features are a major challenge for machine learning methods in general and for deep learning in particular. Existing solutions such as one-hot encoding and entity embeddings can be hard to scale when the cardinality is very high, require much space, are hard to interpret or may overfit the data. A special scenario of interest is that of repeated measures, where the categorical feature is the identity of the individual or object, and each object is measured several times, possibly under different conditions (values of the other features). We propose accounting for high-cardinality categorical features as random effects variables in a regression setting, and consequently adopt the corresponding negative log likelihood loss from the linear mixed models (LMM) statistical literature and integrate it in a deep learning framework. We test our model which we call LMMNN on simulated as well as real datasets with a single categorical feature with high cardinality, using various baseline neural networks architectures such as convolutional networks and LSTM, and various applications in

e-commerce, healthcare and computer vision. Our results show that treating high-cardinality categorical features as random effects leads to a significant improvement in prediction performance compared to state of the art alternatives. Potential extensions such as accounting for multiple categorical features and classification settings are discussed. Our code and simulations are available at https://github.com/gsimchoni/lmmnn.
****
Learning Debiased Representation via Disentangled Feature Augmentation
Jungsoo Lee, Eungyeup Kim, Juyoung Lee, Jihyeon Lee, Jaegul Choo
Image classification models tend to make decisions based on peripheral attributes of data items that have strong correlation with a target variable (i.e., dataset bias). These biased models suffer from the poor generalization capability when evaluated on unbiased datasets. Existing approaches for debiasing often identify and emphasize those samples with no such correlation (i.e., bias-conflicting) without defining the bias type in advance. However, such bias-conflicting samples are significantly scarce in biased datasets, limiting the debiasing capability of these approaches. This paper first presents an empirical analysis revealing that training with "diverse" bias-conflicting samples beyond a given training set is crucial for debiasing as well as the generalization capability. Based on this observation, we propose a novel feature-level data augmentation technique in order to synthesize diverse bias-conflicting samples. To this end, our method learns the disentangled representation of (1) the intrinsic attributes (i.e., those inherently defining a certain class) and (2) bias attributes (i.e., peripheral attributes causing the bias), from a large number of bias-aligned samples, the bias attributes of which have strong correlation with the target variable. Using the disentangled representation, we synthesize bias-conflicting samples that contain the diverse intrinsic attributes of bias-aligned samples by swapping their latent features. By utilizing these diversified bias-conflicting features during the training, our approach achieves superior classification accuracy and debiasing results against the existing baselines on both synthetic and real-world datasets.
****
Scallop: From Probabilistic Deductive Databases to Scalable Differentiable Reasoning
Jiani Huang, Ziyang Li, Binghong Chen, Karan Samel, Mayur Naik, Le Song, Xujie Si
Deep learning and symbolic reasoning are complementary techniques for an intelligent system. However, principled combinations of these techniques have limited scalability, rendering them ill-suited for real-world applications. We propose Scallop, a system that builds upon probabilistic deductive databases, to bridge this gap. The key insight underlying Scallop is a provenance framework that introduces a tunable parameter to specify the level of reasoning granularity. Scallop thereby i) generalizes exact probabilistic reasoning, ii) asymptotically reduces computational cost, and iii) provides relative accuracy guarantees. On a suite of tasks that involve mathematical and logical reasoning, Scallop scales significantly better without sacrificing accuracy compared to DeepProbLog, a principled neural logic programming approach. We also create and evaluate on a real-world Visual Question Answering (VQA) benchmark that requires multi-hop reasoning. Scallop outperforms two VQA-tailored models, a Neural Module Networks based and a transformer based model, by 12.42% and 21.66% respectively.
****
Learning to Synthesize Programs as Interpretable and Generalizable Policies
Dweep Trivedi, Jesse Zhang, Shao-Hua Sun, Joseph J. Lim
Recently, deep reinforcement learning (DRL) methods have achieved impressive performance on tasks in a variety of domains. However, neural network policies produced with DRL methods are not human-interpretable and often have difficulty generalizing to novel scenarios. To address these issues, prior works explore learning programmatic policies that are more interpretable and structured for generalization. Yet, these works either employ limited policy representations (e.g. decision trees, state machines, or predefined program templates) or require stronger

supervision (e.g. input/output state pairs or expert demonstrations). We present a framework that instead learns to synthesize a program, which details the procedure to solve a task in a flexible and expressive manner, solely from reward signals. To alleviate the difficulty of learning to compose programs to induce the desired agent behavior from scratch, we propose to first learn a program embedding space that continuously parameterizes diverse behaviors in an unsupervised manner and then search over the learned program embedding space to yield a program that maximizes the return for a given task. Experimental results demonstrate that the proposed framework not only learns to reliably synthesize task-solving programs but also outperforms DRL and program synthesis baselines while producing interpretable and more generalizable policies. We also justify the necessity of the proposed two-stage learning scheme as well as analyze various methods for learning the program embedding. Website at https://clvrai.com/leaps.
****

The functional specialization of visual cortex emerges from training parallel pathways with self-supervised predictive learning

Shahab Bakhtiari, Patrick Mineault, Timothy Lillicrap, Christopher Pack, Blake Richards

The visual system of mammals is comprised of parallel, hierarchical specialized pathways. Different pathways are specialized in so far as they use representations that are more suitable for supporting specific downstream behaviours. In particular, the clearest example is the specialization of the ventral ("what") and dorsal ("where") pathways of the visual cortex. These two pathways support behaviours related to visual recognition and movement, respectively. To-date, deep neural networks have mostly been used as models of the ventral, recognition pathway. However, it is unknown whether both pathways can be modelled with a single deep ANN. Here, we ask whether a single model with a single loss function can capture the properties of both the ventral and the dorsal pathways. We explore this question using data from mice, who like other mammals, have specialized pathways that appear to support recognition and movement behaviours. We show that when we train a deep neural network architecture with two parallel pathways using a self-supervised predictive loss function, we can outperform other models in fitting mouse visual cortex. Moreover, we can model both the dorsal and ventral pathways. These results demonstrate that a self-supervised predictive learning approach applied to parallel pathway architectures can account for some of the functional specialization seen in mammalian visual systems.
****

Adversarial Training Helps Transfer Learning via Better Representations

Zhun Deng, Linjun Zhang, Kailas Vodrahalli, Kenji Kawaguchi, James Y. Zou

Transfer learning aims to leverage models pre-trained on source data to efficiently adapt to target setting, where only limited data are available for model fine-tuning. Recent works empirically demonstrate that adversarial training in the source data can improve the ability of models to transfer to new domains. However, why this happens is not known. In this paper, we provide a theoretical model to rigorously analyze how adversarial training helps transfer learning. We show that adversarial training in the source data generates provably better representations, so fine-tuning on top of this representation leads to a more accurate predictor of the target data. We further demonstrate both theoretically and empirically that semi-supervised learning in the source data can also improve transfer learning by similarly improving the representation. Moreover, performing adversarial training on top of semi-supervised learning can further improve transferability, suggesting that the two approaches have complementary benefits on representations. We support our theories with experiments on popular data sets and deep learning architectures.
****

Improving Coherence and Consistency in Neural Sequence Models with Dual-System, Neuro-Symbolic Reasoning

Maxwell Nye, Michael Tessler, Josh Tenenbaum, Brenden M. Lake

Human reasoning can be understood as an interplay between two systems: the intuitive and associative ("System 1") and the deliberative and logical ("System 2").

Neural sequence models---which have been increasingly successful at performing complex, structured tasks---exhibit the advantages and failure modes of System 1 : they are fast and learn patterns from data, but are often inconsistent and incoherent. In this work, we seek a lightweight, training-free means of improving existing System 1-like sequence models by adding System 2-inspired logical reasoning. We explore several variations on this theme in which candidate generations from a neural sequence model are examined for logical consistency by a symbolic reasoning module, which can either accept or reject the generations. Our approach uses neural inference to mediate between the neural System 1 and the logical System 2. Results in robust story generation and grounded instruction-following show that this approach can increase the coherence and accuracy of neurally-based generations.
****
Learning the optimal Tikhonov regularizer for inverse problems
Giovanni S. Alberti, Ernesto De Vito, Matti Lassas, Luca Ratti, Matteo Santacesaria
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
NovelD: A Simple yet Effective Exploration Criterion
Tianjun Zhang, Huazhe Xu, Xiaolong Wang, Yi Wu, Kurt Keutzer, Joseph E. Gonzalez, Yuandong Tian
Efficient exploration under sparse rewards remains a key challenge in deep reinforcement learning. Previous exploration methods (e.g., RND) have achieved strong results in multiple hard tasks. However, if there are multiple novel areas to explore, these methods often focus quickly on one without sufficiently trying others (like a depth-wise first search manner). In some scenarios (e.g., four corridor environment in Sec 4.2), we observe they explore in one corridor for long and fail to cover all the states. On the other hand, in theoretical RL, with optimistic initialization and the inverse square root of visitation count as a bonus, it won't suffer from this and explores different novel regions alternatively (like a breadth-first search manner). In this paper, inspired by this, we propose a simple but effective criterion called NovelD by weighting every novel area approximately equally. Our algorithm is very simple but yet shows comparable performance or even outperforms multiple SOTA exploration methods in many hard exploration tasks. Specifically, NovelD solves all the static procedurally-generated tasks in Mini-Grid with just 120M environment steps, without any curriculum learning. In comparison, the previous SOTA only solves 50% of them. NovelD also achieves SOTA on multiple tasks in NetHack, a rogue-like game that contains more challenging procedurally-generated environments. In multiple Atari games (e.g., Monte Zuma's Revenge, Venture, Gravitar), NovelD outperforms RND. We analyze NovelD thoroughly in MiniGrid and found that empirically it helps the agent explore the environment more uniformly with a focus on exploring beyond the boundary.
****
On Margin-Based Cluster Recovery with Oracle 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.
****
Multi-Scale Representation Learning on Proteins
Vignesh Ram Somnath, Charlotte Bunne, Andreas Krause
Proteins are fundamental biological entities mediating key roles in cellular function and disease. This paper introduces a multi-scale graph construction of a protein –HoloProt– connecting surface to structure and sequence. The surface captures coarser details of the protein, while sequence as primary component and structure –comprising secondary and tertiary components– capture finer details. Our

graph encoder then learns a multi-scale representation by allowing each level to integrate the encoding from level(s) below with the graph at that level. We test the learned representation on different tasks, (i.) ligand binding affinity (regression), and (ii.) protein function prediction (classification).On the regression task, contrary to previous methods, our model performs consistently and reliably across different dataset splits, outperforming all baselines on most splits. On the classification task, it achieves a performance close to the top-performing model while using 10x fewer parameters. To improve the memory efficiency of our construction, we segment the multiplex protein surface manifold into molecular superpixels and substitute the surface with these superpixels at little to no performance loss.
****

Sparse Quadratic Optimisation over the Stiefel Manifold with Application to Permutation Synchronisation

Florian Bernard, Daniel Cremers, Johan Thunberg

We address the non-convex optimisation problem of finding a sparse matrix on the Stiefel manifold (matrices with mutually orthogonal columns of unit length) that maximises (or minimises) a quadratic objective function. Optimisation problems on the Stiefel manifold occur for example in spectral relaxations of various combinatorial problems, such as graph matching, clustering, or permutation synchronisation. Although sparsity is a desirable property in such settings, it is mostly neglected in spectral formulations since existing solvers, e.g. based on eigenvalue decomposition, are unable to account for sparsity while at the same time maintaining global optimality guarantees. We fill this gap and propose a simple yet effective sparsity-promoting modification of the Orthogonal Iteration algorithm for finding the dominant eigenspace of a matrix. By doing so, we can guarantee that our method finds a Stiefel matrix that is globally optimal with respect to the quadratic objective function, while in addition being sparse. As a motivating application we consider the task of permutation synchronisation, which can be understood as a constrained clustering problem that has particular relevance for matching multiple images or 3D shapes in computer vision, computer graphics, and beyond. We demonstrate that the proposed approach outperforms previous methods in this domain.
****

Second-Order Neural ODE Optimizer

Guan-Horng Liu, Tianrong Chen, Evangelos Theodorou

We propose a novel second-order optimization framework for training the emerging deep continuous-time models, specifically the Neural Ordinary Differential Equations (Neural ODEs). Since their training already involves expensive gradient computation by solving a backward ODE, deriving efficient second-order methods becomes highly nontrivial. Nevertheless, inspired by the recent Optimal Control (OC) interpretation of training deep networks, we show that a specific continuous-time OC methodology, called Differential Programming, can be adopted to derive backward ODEs for higher-order derivatives at the same O(1) memory cost. We further explore a low-rank representation of the second-order derivatives and show that it leads to efficient preconditioned updates with the aid of Kronecker-based factorization. The resulting method – named SNOpt – converges much faster than first-order baselines in wall-clock time, and the improvement remains consistent across various applications, e.g. image classification, generative flow, and time-series prediction. Our framework also enables direct architecture optimization, such as the integration time of Neural ODEs, with second-order feedback policies, strengthening the OC perspective as a principled tool of analyzing optimization in deep learning. Our code is available at https://github.com/ghliu/snopt.
****

Graph Neural Networks with Local Graph Parameters

Pablo Barceló, Floris Geerts, Juan Reutter, Maksimilian Ryschkov

Various recent proposals increase the distinguishing power of Graph Neural Networks (GNNs) by propagating features between k-tuples of vertices. The distinguishing power of these “higher-order” GNNs is known to be bounded by the k-dimensional Weisfeiler-Leman (WL) test, yet their O(n^k) memory requirements limit their

applicability. Other proposals infuse GNNs with local higher-order graph structu
ral information from the start, hereby inheriting the desirable O(n) memory requ
irement from GNNs at the cost of a one-time, possibly non-linear, preprocessing
step. We propose local graph parameter enabled GNNs as a framework for studying
the latter kind of approaches and precisely characterize their distinguishing po
wer, in terms of a variant of the WL test, and in terms of the graph structural
properties that they can take into account. Local graph parameters can be added
to any GNN architecture, and are cheap to compute. In terms of expressive power,
our proposal lies in the middle of GNNs and their higher-order counterparts. Fu
rther, we propose several techniques to aide in choosing the right local graph p
arameters. Our results connect GNNs with deep results in finite model theory and
finite variable logics. Our experimental evaluation shows that adding local gra
ph parameters often has a positive effect for a variety of GNNs, datasets and gr
aph learning tasks.
****
Closing the Gap: Tighter Analysis of Alternating Stochastic Gradient Methods for
Bilevel Problems
Tianyi Chen, Yuejiao Sun, Wotao Yin
Requests for name changes in the electronic proceedings will be accepted with no
questions asked. However name changes may cause bibliographic tracking issues.
Authors are asked to consider this carefully and discuss it with their co-auth
ors prior to requesting a name change in the electronic proceedings.
****
Dense Unsupervised Learning for Video Segmentation
Nikita Araslanov, Simone Schaub-Meyer, Stefan Roth
We present a novel approach to unsupervised learning for video object segmentati
on (VOS). Unlike previous work, our formulation allows to learn dense feature re
presentations directly in a fully convolutional regime. We rely on uniform grid
sampling to extract a set of anchors and train our model to disambiguate between
them on both inter- and intra-video levels. However, a naive scheme to train su
ch a model results in a degenerate solution. We propose to prevent this with a s
imple regularisation scheme, accommodating the equivariance property of the segm
entation task to similarity transformations. Our training objective admits effic
ient implementation and exhibits fast training convergence. On established VOS b
enchmarks, our approach exceeds the segmentation accuracy of previous work despi
te using significantly less training data and compute power.
****
Charting and Navigating the Space of Solutions for Recurrent Neural Networks
Elia Turner, Kabir V Dabholkar, Omri Barak
In recent years Recurrent Neural Networks (RNNs) were successfully used to model
the way neural activity drives task-related behavior in animals, operating unde
r the implicit assumption that the obtained solutions are universal. Observation
s in both neuroscience and machine learning challenge this assumption. Animals c
an approach a given task with a variety of strategies, and training machine lear
ning algorithms introduces the phenomenon of underspecification. These observati
ons imply that every task is associated with a space of solutions. To date, the
structure of this space is not understood, limiting the approach of comparing RN
Ns with neural data.Here, we characterize the space of solutions associated with
various tasks. We first study a simple two-neuron network on a task that leads
to multiple solutions. We trace the nature of the final solution back to the net
work’s initial connectivity and identify discrete dynamical regimes that underli
e this diversity. We then examine three neuroscience-inspired tasks: Delayed dis
crimination, Interval discrimination, and Time reproduction. For each task, we f
ind a rich set of solutions. One layer of variability can be found directly in t
he neural activity of the networks. An additional layer is uncovered by testing
the trained networks' ability to extrapolate, as a perturbation to a system ofte
n reveals hidden structure. Furthermore, we relate extrapolation patterns to spe
cific dynamical objects and effective algorithms found by the networks. We intro
duce a tool to derive the reduced dynamics of networks by generating a compact d
irected graph describing the essence of the dynamics with regards to behavioral

inputs and outputs. Using this representation, we can partition the solutions to each task into a handful of types and show that neural features can partially predict them.Taken together, our results shed light on the concept of the space of solutions and its uses both in Machine learning and in Neuroscience.
****
Fast Training Method for Stochastic Compositional Optimization Problems
Hongchang Gao, Heng Huang
The stochastic compositional optimization problem covers a wide range of machine learning models, such as sparse additive models and model-agnostic meta-learning. Thus, it is necessary to develop efficient methods for its optimization. Existing methods for the stochastic compositional optimization problem only focus on the single machine scenario, which is far from satisfactory when data are distributed on different devices. To address this problem, we propose novel decentralized stochastic compositional gradient descent methods to efficiently train the large-scale stochastic compositional optimization problem. To the best of our knowledge, our work is the first one facilitating decentralized training for this kind of problem. Furthermore, we provide the convergence analysis for our methods, which shows that the convergence rate of our methods can achieve linear speedup with respect to the number of devices. At last, we apply our decentralized training methods to the model-agnostic meta-learning problem, and the experimental results confirm the superior performance of our methods.
****
Dual-stream Network for Visual Recognition
Mingyuan Mao, peng gao, Renrui Zhang, Honghui Zheng, Teli Ma, Yan Peng, Errui Ding, Baochang Zhang, Shumin Han
Transformers with remarkable global representation capacities achieve competitive results for visual tasks, but fail to consider high-level local pattern information in input images. In this paper, we present a generic Dual-stream Network (DS-Net) to fully explore the representation capacity of local and global pattern features for image classification. Our DS-Net can simultaneously calculate fine-grained and integrated features and efficiently fuse them. Specifically, we propose an Intra-scale Propagation module to process two different resolutions in each block and an Inter-Scale Alignment module to perform information interaction across features at dual scales. Besides, we also design a Dual-stream FPN (DS-FPN) to further enhance contextual information for downstream dense predictions. Without bells and whistles, the proposed DS-Net outperforms DeiT-Small by 2.4\% in terms of top-1 accuracy on ImageNet-1k and achieves state-of-the-art performance over other Vision Transformers and ResNets. For object detection and instance segmentation, DS-Net-Small respectively outperforms ResNet-50 by 6.4\% and 5.5 \% in terms of mAP on MSCOCO 2017, and surpasses the previous state-of-the-art scheme, which significantly demonstrates its potential to be a general backbone in vision tasks. The code will be released soon.
****
Estimating High Order Gradients of the Data Distribution by Denoising
Chenlin Meng, Yang Song, Wenzhe Li, Stefano Ermon
The first order derivative of a data density can be estimated efficiently by denoising score matching, and has become an important component in many applications, such as image generation and audio synthesis. Higher order derivatives provide additional local information about the data distribution and enable new applications. Although they can be estimated via automatic differentiation of a learned density model, this can amplify estimation errors and is expensive in high dimensional settings. To overcome these limitations, we propose a method to directly estimate high order derivatives (scores) of a data density from samples. We first show that denoising score matching can be interpreted as a particular case of Tweedie’s formula. By leveraging Tweedie’s formula on higher order moments, we generalize denoising score matching to estimate higher order derivatives. We demonstrate empirically that models trained with the proposed method can approximate second order derivatives more efficiently and accurately than via automatic differentiation. We show that our models can be used to quantify uncertainty in denoising and to improve the mixing speed of Langevin dynamics via Ozaki discreti

zation for sampling synthetic data and natural images.
****
Machine versus Human Attention in Deep Reinforcement Learning Tasks
Suna (Sihang) Guo, Ruohan Zhang, Bo Liu, Yifeng Zhu, Dana Ballard, Mary Hayhoe, Peter Stone
Deep reinforcement learning (RL) algorithms are powerful tools for solving visuomotor decision tasks. However, the trained models are often difficult to interpret, because they are represented as end-to-end deep neural networks. In this paper, we shed light on the inner workings of such trained models by analyzing the pixels that they attend to during task execution, and comparing them with the pixels attended to by humans executing the same tasks. To this end, we investigate the following two questions that, to the best of our knowledge, have not been previously studied. 1) How similar are the visual representations learned by RL agents and humans when performing the same task? and, 2) How do similarities and differences in these learned representations explain RL agents' performance on these tasks? Specifically, we compare the saliency maps of RL agents against visual attention models of human experts when learning to play Atari games. Further, we analyze how hyperparameters of the deep RL algorithm affect the learned representations and saliency maps of the trained agents. The insights provided have the potential to inform novel algorithms for closing the performance gap between human experts and RL agents.
****
Reusing Combinatorial Structure: Faster Iterative Projections over Submodular Base Polytopes
Jai Moondra, Hassan Mortagy, Swati Gupta
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Constrained Optimization to Train Neural Networks on Critical and Under-Represented Classes
Sara Sangalli, Ertunc Erdil, Andeas Hötker, Olivio Donati, Ender Konukoglu
Deep neural networks (DNNs) are notorious for making more mistakes for the classes that have substantially fewer samples than the others during training. Such class imbalance is ubiquitous in clinical applications and very crucial to handle because the classes with fewer samples most often correspond to critical cases (e.g., cancer) where misclassifications can have severe consequences.Not to miss such cases, binary classifiers need to be operated at high True Positive Rates (TPRs) by setting a higher threshold, but this comes at the cost of very high False Positive Rates (FPRs) for problems with class imbalance. Existing methods for learning under class imbalance most often do not take this into account. We argue that prediction accuracy should be improved by emphasizing the reduction of FPRs at high TPRs for problems where misclassification of the positive, i.e. critical, class samples are associated with higher cost.To this end, we pose the training of a DNN for binary classification as a constrained optimization problem and introduce a novel constraint that can be used with existing loss functions to enforce maximal area under the ROC curve (AUC) through prioritizing FPR reduction at high TPR. We solve the resulting constrained optimization problem using an Augmented Lagrangian method (ALM).Going beyond binary, we also propose two possible extensions of the proposed constraint for multi-class classification problems.We present experimental results for image-based binary and multi-class classification applications using an in-house medical imaging dataset, CIFAR10, and CIFAR100. Our results demonstrate that the proposed method improves the baselines in majority of the cases by attaining higher accuracy on critical classes while reducing the misclassification rate for the non-critical class samples.
****
Collapsed Variational Bounds for Bayesian Neural Networks
Marcin Tomczak, Siddharth Swaroop, Andrew Foong, Richard Turner
Recent interest in learning large variational Bayesian Neural Networks (BNNs) ha

s been partly hampered by poor predictive performance caused by underfitting, and their performance is known to be very sensitive to the prior over weights. Current practice often fixes the prior parameters to standard values or tunes them using heuristics or cross-validation. In this paper, we treat prior parameters in a distributional way by extending the model and collapsing the variational bound with respect to their posteriors. This leads to novel and tighter Evidence Lower Bounds (ELBOs) for performing variational inference (VI) in BNNs. Our experiments show that the new bounds significantly improve the performance of Gaussian mean-field VI applied to BNNs on a variety of data sets, demonstrating that mean-field VI works well even in deep models. We also find that the tighter ELBOs can be good optimization targets for learning the hyperparameters of hierarchical priors.
****
Consistent Estimation for PCA and Sparse Regression with Oblivious Outliers
Tommaso d'Orsi, Chih-Hung Liu, Rajai Nasser, Gleb Novikov, David Steurer, Stefan Tiegel
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Offline Constrained Multi-Objective Reinforcement Learning via Pessimistic Dual Value Iteration
Runzhe Wu, Yufeng Zhang, Zhuoran Yang, Zhaoran Wang
In constrained multi-objective RL, the goal is to learn a policy that achieves the best performance specified by a multi-objective preference function under a constraint. We focus on the offline setting where the RL agent aims to learn the optimal policy from a given dataset. This scenario is common in real-world applications where interactions with the environment are expensive and the constraint violation is dangerous. For such a setting, we transform the original constrained problem into a primal-dual formulation, which is solved via dual gradient ascent. Moreover, we propose to combine such an approach with pessimism to overcome the uncertainty in offline data, which leads to our Pessimistic Dual Iteration (PEDI). We establish upper bounds on both the suboptimality and constraint violation for the policy learned by PEDI based on an arbitrary dataset, which proves that PEDI is provably sample efficient. We also specialize PEDI to the setting with linear function approximation. To the best of our knowledge, we propose the first provably efficient constrained multi-objective RL algorithm with offline data without any assumption on the coverage of the dataset.
****
Absolute Neighbour Difference based Correlation Test for Detecting Heteroscedastic Relationships
Lifeng Zhang
It is a challenge to detect complicated data relationships thoroughly. Here, we propose a new statistical measure, named the absolute neighbour difference based neighbour correlation coefficient, to detect the associations between variables through examining the heteroscedasticity of the unpredictable variation of dependent variables. Different from previous studies, the new method concentrates on measuring nonfunctional relationships rather than functional or mixed associations. Either used alone or in combination with other measures, it enables not only a convenient test of heteroscedasticity, but also measuring functional and nonfunctional relationships separately that obviously leads to a deeper insight into the data associations. The method is concise and easy to implement that does not rely on explicitly estimating the regression residuals or the dependencies between variables so that it is not restrict to any kind of model assumption. The mechanisms of the correlation test are proved in theory and demonstrated with numerical analyses.
****
Batch Multi-Fidelity Bayesian Optimization with Deep Auto-Regressive Networks
Shibo Li, Robert Kirby, Shandian Zhe

Bayesian optimization (BO) is a powerful approach for optimizing black-box, expensive-to-evaluate functions. To enable a flexible trade-off between the cost and accuracy, many applications allow the function to be evaluated at different fidelities. In order to reduce the optimization cost while maximizing the benefit-cost ratio, in this paper we propose Batch Multi-fidelity Bayesian Optimization with Deep Auto-Regressive Networks (BMBO-DARN). We use a set of Bayesian neural networks to construct a fully auto-regressive model, which is expressive enough to capture strong yet complex relationships across all the fidelities, so as to improve the surrogate learning and optimization performance. Furthermore, to enhance the quality and diversity of queries, we develop a simple yet efficient batch querying method, without any combinatorial search over the fidelities. We propose a batch acquisition function based on Max-value Entropy Search (MES) principle, which penalizes highly correlated queries and encourages diversity. We use posterior samples and moment matching to fulfill efficient computation of the acquisition function, and conduct alternating optimization over every fidelity-input pair, which guarantees an improvement at each step. We demonstrate the advantage of our approach on four real-world hyperparameter optimization applications.
****
Mastering Atari Games with Limited Data
Weirui Ye, Shaohuai Liu, Thanard Kurutach, Pieter Abbeel, Yang Gao
Reinforcement learning has achieved great success in many applications. However, sample efficiency remains a key challenge, with prominent methods requiring millions (or even billions) of environment steps to train. Recently, there has been significant progress in sample efficient image-based RL algorithms; however, consistent human-level performance on the Atari game benchmark remains an elusive goal. We propose a sample efficient model-based visual RL algorithm built on MuZero, which we name EfficientZero. Our method achieves 194.3% mean human performance and 109.0% median performance on the Atari 100k benchmark with only two hours of real-time game experience and outperforms the state SAC in some tasks on the DMControl 100k benchmark. This is the first time an algorithm achieves super-human performance on Atari games with such little data. EfficientZero's performance is also close to DQN's performance at 200 million frames while we consume 500 times less data. EfficientZero's low sample complexity and high performance can bring RL closer to real-world applicability. We implement our algorithm in an easy-to-understand manner and it is available at https://github.com/YeWR/EfficientZero. We hope it will accelerate the research of MCTS-based RL algorithms in the wider community.
****
Dealing With Misspecification In Fixed-Confidence Linear Top-m Identification
Clémence Réda, Andrea Tirinzoni, Rémy Degenne
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 Generalization in RL is Difficult: Epistemic POMDPs and Implicit Partial Observability
Dibya Ghosh, Jad Rahme, Aviral Kumar, Amy Zhang, Ryan P. Adams, Sergey Levine
Generalization is a central challenge for the deployment of reinforcement learning (RL) systems in the real world. In this paper, we show that the sequential structure of the RL problem necessitates new approaches to generalization beyond the well-studied techniques used in supervised learning. While supervised learning methods can generalize effectively without explicitly accounting for epistemic uncertainty, we describe why appropriate uncertainty handling can actually be essential in RL. We show that generalization to unseen test conditions from a limited number of training conditions induces a kind of implicit partial observability, effectively turning even fully-observed MDPs into POMDPs. Informed by this observation, we recast the problem of generalization in RL as solving the induced partially observed Markov decision process, which we call the epistemic POMDP.

We demonstrate the failure modes of algorithms that do not appropriately handle this partial observability, and suggest a simple ensemble-based technique for approximately solving the partially observed problem. Empirically, we demonstrate that our simple algorithm derived from the epistemic POMDP achieves significant gains in generalization over current methods on the Procgen benchmark suite.
****
Set Prediction in the Latent Space
Konpat Preechakul, Chawan Piansaddhayanon, Burin Naowarat, Tirasan Khandhawit, Sira Sriswasdi, Ekapol Chuangsuwanich
Set prediction tasks require the matching between predicted set and ground truth set in order to propagate the gradient signal. Recent works have performed this matching in the original feature space thus requiring predefined distance functions. We propose a method for learning the distance function by performing the matching in the latent space learned from encoding networks. This method enables the use of teacher forcing which was not possible previously since matching in the feature space must be computed after the entire output sequence is generated. Nonetheless, a naive implementation of latent set prediction might not converge due to permutation instability. To address this problem, we provide sufficient conditions for permutation stability which begets an algorithm to improve the overall model convergence. Experiments on several set prediction tasks, including image captioning and object detection, demonstrate the effectiveness of our method.
****
Best of Both Worlds: Practical and Theoretically Optimal Submodular Maximization in Parallel
Yixin Chen, Tonmoy Dey, Alan Kuhnle
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Fine-grained Generalization Analysis of Inductive Matrix Completion
Antoine Ledent, Rodrigo Alves, Yunwen Lei, 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-authors prior to requesting a name change in the electronic proceedings.
****
Learning Frequency Domain Approximation for Binary Neural Networks
Yixing Xu, Kai Han, Chang Xu, Yehui Tang, Chunjing XU, Yunhe Wang
Binary neural networks (BNNs) represent original full-precision weights and activations into 1-bit with sign function. Since the gradient of the conventional sign function is almost zero everywhere which cannot be used for back-propagation, several attempts have been proposed to alleviate the optimization difficulty by using approximate gradient. However, those approximations corrupt the main direction of factual gradient. To this end, we propose to estimate the gradient of sign function in the Fourier frequency domain using the combination of sine functions for training BNNs, namely frequency domain approximation (FDA). The proposed approach does not affect the low-frequency information of the original sign function which occupies most of the overall energy, and high-frequency coefficients will be ignored to avoid the huge computational overhead. In addition, we embed a noise adaptation module into the training phase to compensate the approximation error. The experiments on several benchmark datasets and neural architectures illustrate that the binary network learned using our method achieves the state-of-the-art accuracy. Code will be available at https://gitee.com/mindspore/models/tree/master/research/cv/FDA-BNN.
****
Reformulating Zero-shot Action Recognition for Multi-label Actions
Alec Kerrigan, Kevin Duarte, Yogesh Rawat, Mubarak Shah
The goal of zero-shot action recognition (ZSAR) is to classify action classes wh

ich were not previously seen during training. Traditionally, this is achieved by training a network to map, or regress, visual inputs to a semantic space where a nearest neighbor classifier is used to select the closest target class. We argue that this approach is sub-optimal due to the use of nearest neighbor on static semantic space and is ineffective when faced with multi-label videos - where two semantically distinct co-occurring action categories cannot be predicted with high confidence. To overcome these limitations, we propose a ZSAR framework which does not rely on nearest neighbor classification, but rather consists of a pairwise scoring function. Given a video and a set of action classes, our method predicts a set of confidence scores for each class independently. This allows for the prediction of several semantically distinct classes within one video input. Our evaluations show that our method not only achieves strong performance on three single-label action classification datasets (UCF-101, HMDB, and RareAct), but also outperforms previous ZSAR approaches on a challenging multi-label dataset (AVA) and a real-world surprise activity detection dataset (MEVA).
****
Optimal Best-Arm Identification Methods for Tail-Risk Measures
Shubhada Agrawal, Wouter M. Koolen, Sandeep Juneja
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
SyMetric: Measuring the Quality of Learnt Hamiltonian Dynamics Inferred from Vision
Irina Higgins, Peter Wirnsberger, Andrew Jaegle, Aleksandar Botev
A recently proposed class of models attempts to learn latent dynamics from high-dimensional observations, like images, using priors informed by Hamiltonian mechanics. While these models have important potential applications in areas like robotics or autonomous driving, there is currently no good way to evaluate their performance: existing methods primarily rely on image reconstruction quality, which does not always reflect the quality of the learnt latent dynamics. In this work, we empirically highlight the problems with the existing measures and develop a set of new measures, including a binary indicator of whether the underlying Hamiltonian dynamics have been faithfully captured, which we call Symplecticity Metric or SyMetric. Our measures take advantage of the known properties of Hamiltonian dynamics and are more discriminative of the model's ability to capture the underlying dynamics than reconstruction error. Using SyMetric, we identify a set of architectural choices that significantly improve the performance of a previously proposed model for inferring latent dynamics from pixels, the Hamiltonian Generative Network (HGN). Unlike the original HGN, the new SyMetric is able to discover an interpretable phase space with physically meaningful latents on some datasets. Furthermore, it is stable for significantly longer rollouts on a diverse range of 13 datasets, producing rollouts of essentially infinite length both forward and backwards in time with no degradation in quality on a subset of the datasets.
****
Learning with Holographic Reduced Representations
Ashwinkumar Ganesan, Hang Gao, Sunil Gandhi, Edward Raff, Tim Oates, James Holt, Mark McLean
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 Barrier Certificates: Towards Safe Reinforcement Learning with Zero Training-time Violations
Yuping Luo, Tengyu Ma
Training-time safety violations have been a major concern when we deploy reinforcement learning algorithms in the real world.This paper explores the possibility

of safe RL algorithms with zero training-time safety violations in the challenging setting where we are only given a safe but trivial-reward initial policy without any prior knowledge of the dynamics and additional offline data.We propose an algorithm, Co-trained Barrier Certificate for Safe RL (CRABS), which iteratively learns barrier certificates, dynamics models, and policies. The barrier certificates are learned via adversarial training and ensure the policy's safety assuming calibrated learned dynamics. We also add a regularization term to encourage larger certified regions to enable better exploration. Empirical simulations show that zero safety violations are already challenging for a suite of simple environments with only 2-4 dimensional state space, especially if high-reward policies have to visit regions near the safety boundary. Prior methods require hundreds of violations to achieve decent rewards on these tasks, whereas our proposed algorithms incur zero violations.
****

On the Second-order Convergence Properties of Random Search Methods
Aurelien Lucchi, Antonio Orvieto, Adamos Solomou
We study the theoretical convergence properties of random-search methods when optimizing non-convex objective functions without having access to derivatives. We prove that standard random-search methods that do not rely on second-order information converge to a second-order stationary point. However, they suffer from an exponential complexity in terms of the input dimension of the problem. In order to address this issue, we propose a novel variant of random search that exploits negative curvature by only relying on function evaluations. We prove that this approach converges to a second-order stationary point at a much faster rate than vanilla methods: namely, the complexity in terms of the number of function evaluations is only linear in the problem dimension. We test our algorithm empirically and find good agreements with our theoretical results.
****

Noether’s Learning Dynamics: Role of Symmetry Breaking in Neural Networks
Hidenori Tanaka, Daniel Kunin
In nature, symmetry governs regularities, while symmetry breaking brings texture. In artificial neural networks, symmetry has been a central design principle to efficiently capture regularities in the world, but the role of symmetry breaking is not well understood. Here, we develop a theoretical framework to study the "geometry of learning dynamics" in neural networks, and reveal a key mechanism of explicit symmetry breaking behind the efficiency and stability of modern neural networks. To build this understanding, we model the discrete learning dynamics of gradient descent using a continuous-time Lagrangian formulation, in which the learning rule corresponds to the kinetic energy and the loss function corresponds to the potential energy. Then, we identify "kinetic symmetry breaking" (KSB), the condition when the kinetic energy explicitly breaks the symmetry of the potential function. We generalize Noether’s theorem known in physics to take into account KSB and derive the resulting motion of the Noether charge: "Noether's Learning Dynamics" (NLD). Finally, we apply NLD to neural networks with normalization layers and reveal how KSB introduces a mechanism of implicit adaptive optimization, establishing an analogy between learning dynamics induced by normalization layers and RMSProp. Overall, through the lens of Lagrangian mechanics, we have established a theoretical foundation to discover geometric design principles for the learning dynamics of neural networks.
****

A Theory of the Distortion-Perception Tradeoff in Wasserstein Space
Dror Freirich, Tomer Michaeli, Ron Meir
The lower the distortion of an estimator, the more the distribution of its outputs generally deviates from the distribution of the signals it attempts to estimate. This phenomenon, known as the perception-distortion tradeoff, has captured significant attention in image restoration, where it implies that fidelity to ground truth images comes on the expense of perceptual quality (deviation from statistics of natural images). However, despite the increasing popularity of performing comparisons on the perception-distortion plane, there remains an important open question: what is the minimal distortion that can be achieved under a given

perception constraint? In this paper, we derive a closed form expression for thi s distortion-perception (DP) function for the mean squared-error (MSE) distortio n and Wasserstein-2 perception index. We prove that the DP function is always qu adratic, regardless of the underlying distribution. This stems from the fact tha t estimators on the DP curve form a geodesic in Wasserstein space. In the Gaussi an setting, we further provide a closed form expression for such estimators. For general distributions, we show how these estimators can be constructed from the estimators at the two extremes of the tradeoff: The global MSE minimizer, and a minimizer of the MSE under a perfect perceptual quality constraint. The latter can be obtained as a stochastic transformation of the former.
****

Neural Production Systems
Anirudh Goyal ALIAS PARTH GOYAL, Aniket Didolkar, Nan Rosemary Ke, Charles Blund ell, Philippe Beaudoin, Nicolas Heess, Michael C. Mozer, Yoshua Bengio
Visual environments are structured, consisting of distinct objects or entities. These entities have properties---visible or latent---that determine the manner in which they interact with one another. To partition images into entities, deep -learning researchers have proposed structural inductive biases such as slot-bas ed architectures. To model interactions among entities, equivariant graph neural nets (GNNs) are used, but these are not particularly well suited to the task fo r two reasons. First, GNNs do not predispose interactions to be sparse, as relat ionships among independent entities are likely to be. Second, GNNs do not facto rize knowledge about interactions in an entity-conditional manner. As an altern ative, we take inspiration from cognitive science and resurrect a classic approa ch, production systems, which consist of a set of rule templates that are applie d by binding placeholder variables in the rules to specific entities. Rules are scored on their match to entities, and the best fitting rules are applied to up date entity properties. In a series of experiments, we demonstrate that this arc hitecture achieves a flexible, dynamic flow of control and serves to factorize e ntity-specific and rule-based information. This disentangling of knowledge achie ves robust future-state prediction in rich visual environments, outperforming st ate-of-the-art methods using GNNs, and allows for the extrapolation from simple (few object) environments to more complex environments.
****

Smoothness Matrices Beat Smoothness Constants: Better Communication Compression Techniques for Distributed Optimization
Mher Safaryan, Filip Hanzely, Peter Richtarik
Large scale distributed optimization has become the default tool for the trainin g of supervised machine learning models with a large number of parameters and tr aining data. Recent advancements in the field provide several mechanisms for spe eding up the training, including {\em compressed communication}, {\em variance r eduction} and {\em acceleration}. However, none of these methods is capable of e xploiting the inherently rich data-dependent smoothness structure of the local l osses beyond standard smoothness constants. In this paper, we argue that when tr aining supervised models, {\em smoothness matrices}---information-rich generali zations of the ubiquitous smoothness constants---can and should be exploited for further dramatic gains, both in theory and practice. In order to further allevi ate the communication burden inherent in distributed optimization, we propose a novel communication sparsification strategy that can take full advantage of the smoothness matrices associated with local losses. To showcase the power of this tool, we describe how our sparsification technique can be adapted to three distr ibuted optimization algorithms---DCGD, DIANA and ADIANA---yielding significant s avings in terms of communication complexity. The new methods always outperform the baselines, often dramatically so.
****

Increasing Liquid State Machine Performance with Edge-of-Chaos Dynamics Organize d by Astrocyte-modulated Plasticity
Vladimir Ivanov, Konstantinos Michmizos
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

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 Sortition Made Transparent
Bailey Flanigan, Gregory Kehne, Ariel D. Procaccia
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 Max-Min Entropy Framework for Reinforcement Learning
Seungyul Han, Youngchul Sung
In this paper, we propose a max-min entropy framework for reinforcement learning (RL) to overcome the limitation of the soft actor-critic (SAC) algorithm implementing the maximum entropy RL in model-free sample-based learning. Whereas the maximum entropy RL guides learning for policies to reach states with high entropy in the future, the proposed max-min entropy framework aims to learn to visit states with low entropy and maximize the entropy of these low-entropy states to promote better exploration. For general Markov decision processes (MDPs), an efficient algorithm is constructed under the proposed max-min entropy framework based on disentanglement of exploration and exploitation. Numerical results show that the proposed algorithm yields drastic performance improvement over the current state-of-the-art RL algorithms.
****
Reward is enough for convex MDPs
Tom Zahavy, Brendan O'Donoghue, Guillaume Desjardins, Satinder Singh
Maximising a cumulative reward function that is Markov and stationary, i.e., defined over state-action pairs and independent of time, is sufficient to capture many kinds of goals in a Markov decision process (MDP). However, not all goals can be captured in this manner. In this paper we study convex MDPs in which goals are expressed as convex functions of the stationary distribution and show that they cannot be formulated using stationary reward functions. Convex MDPs generalize the standard reinforcement learning (RL) problem formulation to a larger framework that includes many supervised and unsupervised RL problems, such as apprenticeship learning, constrained MDPs, and so-called pure exploration'. Our approach is to reformulate the convex MDP problem as a min-max game involving policy and cost (negative reward)players', using Fenchel duality. We propose a meta-algorithm for solving this problem and show that it unifies many existing algorithms in the literature.
****
Fast Doubly-Adaptive MCMC to Estimate the Gibbs Partition Function with Weak Mixing Time Bounds
Shahrzad Haddadan, Yue Zhuang, Cyrus Cousins, Eli Upfal
We present a novel method for reducing the computational complexity of rigorously estimating the partition functions of Gibbs (or Boltzmann) distributions, which arise ubiquitously in probabilistic graphical models. A major obstacle to applying the Gibbs distribution in practice is the need to estimate their partition function (normalizing constant). The state of the art in addressing this problem is multi-stage algorithms which consist of a cooling schedule and a mean estimator in each step of the schedule. While the cooling schedule in these algorithms is adaptive, the mean estimate computations use MCMC as a black-box to draw approximately-independent samples. Here we develop a doubly adaptive approach, combining the adaptive cooling schedule with an adaptive MCMC mean estimator, whose number of Markov chain steps adapts dynamically to the underlying chain. Through rigorous theoretical analysis, we prove that our method outperforms the state of the art algorithms in several factors: (1) The computational complexity of our method is smaller; (2) Our method is less sensitive to loose bounds on mixing times, an inherent components in these algorithms; and (3) The improvement obtained by our method is particularly significant in the most challenging regime of high precision estimates. We demonstrate the advantage of our method in experim

ents run on classic factor graphs, such as voting models and Ising models.
****
Does enforcing fairness mitigate biases caused by subpopulation shift?
Subha Maity, Debarghya Mukherjee, Mikhail Yurochkin, Yuekai Sun
Many instances of algorithmic bias are caused by subpopulation shifts. For example, ML models often perform worse on demographic groups that are underrepresented in the training data. In this paper, we study whether enforcing algorithmic fairness during training improves the performance of the trained model in the \emph{target domain}. On one hand, we conceive scenarios in which enforcing fairness does not improve performance in the target domain. In fact, it may even harm performance. On the other hand, we derive necessary and sufficient conditions under which enforcing algorithmic fairness leads to the Bayes model in the target domain. We also illustrate the practical implications of our theoretical results in simulations and on real data.
****
Implicit Deep Adaptive Design: Policy-Based Experimental Design without Likelihoods
Desi R Ivanova, Adam Foster, Steven Kleinegesse, Michael U. Gutmann, Thomas Rainforth
We introduce implicit Deep Adaptive Design (iDAD), a new method for performing adaptive experiments in real-time with implicit models. iDAD amortizes the cost of Bayesian optimal experimental design (BOED) by learning a design policy network upfront, which can then be deployed quickly at the time of the experiment. The iDAD network can be trained on any model which simulates differentiable samples, unlike previous design policy work that requires a closed form likelihood and conditionally independent experiments. At deployment, iDAD allows design decisions to be made in milliseconds, in contrast to traditional BOED approaches that require heavy computation during the experiment itself. We illustrate the applicability of iDAD on a number of experiments, and show that it provides a fast and effective mechanism for performing adaptive design with implicit models.
****
Sample-Efficient Learning of Stackelberg Equilibria in General-Sum Games
Yu Bai, Chi Jin, Huan Wang, Caiming Xiong
Real world applications such as economics and policy making often involve solving multi-agent games with two unique features: (1) The agents are inherently asymmetric and partitioned into leaders and followers; (2) The agents have different reward functions, thus the game is general-sum. The majority of existing results in this field focuses on either symmetric solution concepts (e.g. Nash equilibrium) or zero-sum games. It remains open how to learn the Stackelberg equilibrium---an asymmetric analog of the Nash equilibrium---in general-sum games efficiently from noisy samples. This paper initiates the theoretical study of sample-efficient learning of the Stackelberg equilibrium, in the bandit feedback setting where we only observe noisy samples of the reward. We consider three representative two-player general-sum games: bandit games, bandit-reinforcement learning (bandit-RL) games, and linear bandit games. In all these games, we identify a fundamental gap between the exact value of the Stackelberg equilibrium and its estimated version using finitely many noisy samples, which can not be closed information-theoretically regardless of the algorithm. We then establish sharp positive results on sample-efficient learning of Stackelberg equilibrium with value optimal up to the gap identified above, with matching lower bounds in the dependency on the gap, error tolerance, and the size of the action spaces. Overall, our results unveil unique challenges in learning Stackelberg equilibria under noisy bandit feedback, which we hope could shed light on future research on this topic.
****
Non-approximate Inference for Collective Graphical Models on Path Graphs via Discrete Difference of Convex Algorithm
Yasunori Akagi, Naoki Marumo, Hideaki Kim, Takeshi Kurashima, Hiroyuki Toda
The importance of aggregated count data, which is calculated from the data of multiple individuals, continues to increase. Collective Graphical Model (CGM) is a probabilistic approach to the analysis of aggregated data. One of the most impo

rtant operations in CGM is maximum a posteriori (MAP) inference of unobserved variables under given observations. Because the MAP inference problem for general CGMs has been shown to be NP-hard, an approach that solves an approximate problem has been proposed. However, this approach has two major drawbacks. First, the quality of the solution deteriorates when the values in the count tables are small, because the approximation becomes inaccurate. Second, since continuous relaxation is applied, the integrality constraints of the output are violated. To resolve these problems, this paper proposes a new method for MAP inference for CGMs on path graphs. Our method is based on the Difference of Convex Algorithm (DCA), which is a general methodology to minimize a function represented as the sum of a convex function and a concave function. In our algorithm, important subroutines in DCA can be efficiently calculated by minimum convex cost flow algorithms. Experiments show that the proposed method outputs higher quality solutions than the conventional approach.
****
Implicit Task-Driven Probability Discrepancy Measure for Unsupervised Domain Adaptation
Mao Li, Kaiqi Jiang, Xinhua Zhang
Probability discrepancy measure is a fundamental construct for numerous machine learning models such as weakly supervised learning and generative modeling. However, most measures overlook the fact that the distributions are not the end-product of learning, but are the basis of downstream predictor. Therefore it is important to warp the probability discrepancy measure towards the end tasks, and we hence propose a new bi-level optimization based approach so that the two distributions are compared not uniformly against the entire hypothesis space, but only with respect to the optimal predictor for the downstream end task. When applied to margin disparity discrepancy and contrastive domain discrepancy, our method significantly improves the performance in unsupervised domain adaptation, and enjoys a much more principled training process.
****
SBO-RNN: Reformulating Recurrent Neural Networks via Stochastic Bilevel Optimization
Ziming Zhang, Yun Yue, Guojun Wu, Yanhua Li, Haichong Zhang
In this paper we consider the training stability of recurrent neural networks (RNNs) and propose a family of RNNs, namely SBO-RNN, that can be formulated using stochastic bilevel optimization (SBO). With the help of stochastic gradient descent (SGD), we manage to convert the SBO problem into an RNN where the feedforward and backpropagation solve the lower and upper-level optimization for learning hidden states and their hyperparameters, respectively. We prove that under mild conditions there is no vanishing or exploding gradient in training SBO-RNN. Empirically we demonstrate our approach with superior performance on several benchmark datasets, with fewer parameters, less training data, and much faster convergence. Code is available at https://zhang-vislab.github.io.
****
Navigating to the Best Policy in Markov Decision Processes
Aymen Al Marjani, Aurélien Garivier, 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-authors prior to requesting a name change in the electronic proceedings.
****
A Faster Decentralized Algorithm for Nonconvex Minimax Problems
Wenhan Xian, Feihu Huang, Yanfu Zhang, Heng Huang
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Generalization Bounds For Meta-Learning: An Information-Theoretic Analysis
Qi CHEN, Changjian Shui, Mario Marchand

We derive a novel information-theoretic analysis of the generalization property
of meta-learning algorithms. Concretely, our analysis proposes a generic underst
anding in both the conventional learning-to-learn framework \citep{amit2018meta}
 and the modern model-agnostic meta-learning (MAML) algorithms \citep{finn2017mo
del}.Moreover, we provide a data-dependent generalization bound for the stochast
ic variant of MAML, which is \emph{non-vacuous} for deep few-shot learning. As c
ompared to previous bounds that depend on the square norms of gradients, empiric
al validations on both simulated data and a well-known few-shot benchmark show t
hat our bound is orders of magnitude tighter in most conditions.
****
ReLU Regression with Massart Noise
Ilias Diakonikolas, Jong Ho Park, Christos Tzamos
Requests for name changes in the electronic proceedings will be accepted with no
 questions asked. However name changes may cause bibliographic tracking issues.
  Authors are asked to consider this carefully and discuss it with their co-auth
ors prior to requesting a name change in the electronic proceedings.
****
Identification of the Generalized Condorcet Winner in Multi-dueling Bandits
Björn Haddenhorst, Viktor Bengs, Eyke Hüllermeier
The reliable identification of the “best” arm while keeping the sample complexit
y as low as possible is a common task in the field of multi-armed bandits. In th
e multi-dueling variant of multi-armed bandits, where feedback is provided in th
e form of a winning arm among as set of k chosen ones, a reasonable notion of be
st arm is the generalized Condorcet winner (GCW). The latter is an the arm that
has the greatest probability of being the winner in each subset containing it. I
n this paper, we derive lower bounds on the sample complexity for the task of id
entifying the GCW under various assumptions. As a by-product, our lower bound re
sults provide new insights for the special case of dueling bandits (k = 2). We p
ropose the Dvoretzky–Kiefer–Wolfowitz tournament (DKWT) algorithm, which we prov
e to be nearly optimal. In a numerical study, we show that DKWT empirically outp
erforms current state-of-the-art algorithms, even in the special case of dueling
 bandits or under a Plackett-Luce assumption on the feedback mechanism.
****
Robust Inverse Reinforcement Learning under Transition Dynamics Mismatch
Luca Viano, Yu-Ting Huang, Parameswaran Kamalaruban, Adrian Weller, Volkan Cevhe
r
Requests for name changes in the electronic proceedings will be accepted with 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.
****
Re-ranking for image retrieval and transductive few-shot classification
Xi SHEN, Yang Xiao, Shell Xu Hu, Othman Sbai, Mathieu Aubry
In the problems of image retrieval and few-shot classification, the mainstream a
pproaches focus on learning a better feature representation. However, directly t
ackling the distance or similarity measure between images could also be efficien
t. To this end, we revisit the idea of re-ranking the top-k retrieved images in
the context of image retrieval (e.g., the k-reciprocal nearest neighbors) and ge
neralize this idea to transductive few-shot learning. We propose to meta-learn t
he re-ranking updates such that the similarity graph converges towards the targe
t similarity graph induced by the image labels. Specifically, the re-ranking mod
ule takes as input an initial similarity graph between the query image and the c
ontextual images using a pre-trained feature extractor, and predicts an improved
 similarity graph by leveraging the structure among the involved images. We show
 that our re-ranking approach can be applied to unseen images and can further bo
ost existing approaches for both image retrieval and few-shot learning problems.
 Our approach operates either independently or in conjunction with classical re-
ranking approaches, yielding clear and consistent improvements on image retrieva
l (CUB, Cars, SOP, rOxford5K and rParis6K) and transductive few-shot classificat
ion (Mini-ImageNet, tiered-ImageNet and CIFAR-FS) benchmarks. Our code is availa

ble at https://imagine.enpc.fr/~shenx/SSR/.
****
Post-processing for Individual Fairness
Felix Petersen, Debarghya Mukherjee, Yuekai Sun, Mikhail Yurochkin
Post-processing in algorithmic fairness is a versatile approach for correcting bias in ML systems that are already used in production. The main appeal of post-processing is that it avoids expensive retraining. In this work, we propose general post-processing algorithms for individual fairness (IF). We consider a setting where the learner only has access to the predictions of the original model and a similarity graph between individuals, guiding the desired fairness constraints. We cast the IF post-processing problem as a graph smoothing problem corresponding to graph Laplacian regularization that preserves the desired "treat similar individuals similarly" interpretation. Our theoretical results demonstrate the connection of the new objective function to a local relaxation of the original individual fairness. Empirically, our post-processing algorithms correct individual biases in large-scale NLP models such as BERT, while preserving accuracy.
****
OpenMatch: Open-Set Semi-supervised Learning with Open-set Consistency Regularization
Kuniaki Saito, Donghyun Kim, Kate Saenko
Semi-supervised learning (SSL) is an effective means to leverage unlabeled data to improve a model’s performance. Typical SSL methods like FixMatch assume that labeled and unlabeled data share the same label space. However, in practice, unlabeled data can contain categories unseen in the labeled set, i.e., outliers, which can significantly harm the performance of SSL algorithms. To address this problem, we propose a novel Open-set Semi-Supervised Learning (OSSL) approach called OpenMatch.Learning representations of inliers while rejecting outliers is essential for the success of OSSL. To this end, OpenMatch unifies FixMatch with novelty detection based on one-vs-all (OVA) classifiers. The OVA-classifier outputs the confidence score of a sample being an inlier, providing a threshold to detect outliers. Another key contribution is an open-set soft-consistency regularization loss, which enhances the smoothness of the OVA-classifier with respect to input transformations and greatly improves outlier detection. \ours achieves state-of-the-art performance on three datasets, and even outperforms a fully supervised model in detecting outliers unseen in unlabeled data on CIFAR10. The code is available at \url{https://github.com/VisionLearningGroup/OP_Match}.
****
End-to-End Training of Multi-Document Reader and Retriever for Open-Domain Question Answering
Devendra Singh, Siva Reddy, Will Hamilton, Chris Dyer, Dani Yogatama
We present an end-to-end differentiable training method for retrieval-augmented open-domain question answering systems that combine information from multiple retrieved documents when generating answers. We model retrieval decisions as latent variables over sets of relevant documents. Since marginalizing over sets of retrieved documents is computationally hard, we approximate this using an expectation-maximization algorithm. We iteratively estimate the value of our latent variable (the set of relevant documents for a given question) and then use this estimate to update the retriever and reader parameters. We hypothesize that such end-to-end training allows training signals to flow to the reader and then to the retriever better than staged-wise training. This results in a retriever that is able to select more relevant documents for a question and a reader that is trained on more accurate documents to generate an answer. Experiments on three benchmark datasets demonstrate that our proposed method outperforms all existing approaches of comparable size by 2-3% absolute exact match points, achieving new state-of-the-art results. Our results also demonstrate the feasibility of learning to retrieve to improve answer generation without explicit supervision of retrieval decisions.
****
Fast Algorithms for L_∞-constrained S-rectangular Robust MDPs
Bahram Behzadian, Marek Petrik, Chin Pang Ho

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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-optimal Mean Estimation Under Differential Privacy
Ziyue Huang, Yuting Liang, Ke Yi
Mean estimation under differential privacy is a fundamental problem, but worst-case optimal mechanisms do not offer meaningful utility guarantees in practice when the global sensitivity is very large. Instead, various heuristics have been proposed to reduce the error on real-world data that do not resemble the worst-case instance. This paper takes a principled approach, yielding a mechanism that is instance-optimal in a strong sense. In addition to its theoretical optimality, the mechanism is also simple and practical, and adapts to a variety of data characteristics without the need of parameter tuning. It easily extends to the local and shuffle model as well.
****
Look at the Variance! Efficient Black-box Explanations with Sobol-based Sensitivity Analysis
Thomas FEL, Remi Cadene, Mathieu Chalvidal, Matthieu Cord, David Vigouroux, Thomas Serre
We describe a novel attribution method which is grounded in Sensitivity Analysis and uses Sobol indices. Beyond modeling the individual contributions of image regions, Sobol indices provide an efficient way to capture higher-order interactions between image regions and their contributions to a neural network's prediction through the lens of variance.We describe an approach that makes the computation of these indices efficient for high-dimensional problems by using perturbation masks coupled with efficient estimators to handle the high dimensionality of images.Importantly, we show that the proposed method leads to favorable scores on standard benchmarks for vision (and language models) while drastically reducing the computing time compared to other black-box methods -- even surpassing the accuracy of state-of-the-art white-box methods which require access to internal representations. Our code is freely available:github.com/fel-thomas/Sobol-Attribution-Method.
****
PatchGame: Learning to Signal Mid-level Patches in Referential Games
Kamal Gupta, Gowthami Somepalli, Anubhav Anubhav, Vinoj Yasanga Jayasundara Magalle Hewa, Matthias Zwicker, Abhinav Shrivastava
We study a referential game (a type of signaling game) where two agents communicate with each other via a discrete bottleneck to achieve a common goal. In our referential game, the goal of the speaker is to compose a message or a symbolic representation of "important" image patches, while the task for the listener is to match the speaker's message to a different view of the same image. We show that it is indeed possible for the two agents to develop a communication protocol without explicit or implicit supervision. We further investigate the developed protocol and show the applications in speeding up recent Vision Transformers by using only important patches, and as pre-training for downstream recognition tasks (e.g., classification).
****
Implicit Generative Copulas
Tim Janke, Mohamed Ghanmi, Florian Steinke
Copulas are a powerful tool for modeling multivariate distributions as they allow to separately estimate the univariate marginal distributions and the joint dependency structure. However, known parametric copulas offer limited flexibility especially in high dimensions, while commonly used non-parametric methods suffer from the curse of dimensionality. A popular remedy is to construct a tree-based hierarchy of conditional bivariate copulas.In this paper, we propose a flexible, yet conceptually simple alternative based on implicit generative neural networks.The key challenge is to ensure marginal uniformity of the estimated copula distribution.We achieve this by learning a multivariate latent distribution with un

specified marginals but the desired dependency structure.By applying the probability integral transform, we can then obtain samples from the high-dimensional copula distribution without relying on parametric assumptions or the need to find a suitable tree structure.Experiments on synthetic and real data from finance, physics, and image generation demonstrate the performance of this approach.
****
Tensor Normal Training for Deep Learning Models
Yi Ren, Donald Goldfarb
Despite the predominant use of first-order methods for training deep learning models, second-order methods, and in particular, natural gradient methods, remain of interest because of their potential for accelerating training through the use of curvature information. Several methods with non-diagonal preconditioning matrices, including KFAC, Shampoo, and K-BFGS, have been proposed and shown to be effective. Based on the so-called tensor normal (TN) distribution, we propose and analyze a brand new approximate natural gradient method, Tensor Normal Training (TNT), which like Shampoo, only requires knowledge of the shape of the training parameters. By approximating the probabilistically based Fisher matrix, as opposed to the empirical Fisher matrix, our method uses the block-wise covariance of the sampling based gradient as the pre-conditioning matrix. Moreover, the assumption that the sampling-based (tensor) gradient follows a TN distribution, ensures that its covariance has a Kronecker separable structure, which leads to a tractable approximation to the Fisher matrix. Consequently, TNT's memory requirements and per-iteration computational costs are only slightly higher than those for first-order methods. In our experiments, TNT exhibited superior optimization performance to state-of-the-art first-order methods, and comparable optimization performance to the state-of-the-art second-order methods KFAC and Shampoo. Moreover, TNT demonstrated its ability to generalize as well as first-order methods, while using fewer epochs.
****
Unintended Selection: Persistent Qualification Rate Disparities and Interventions
Reilly Raab, Yang 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.
****
Revisiting 3D Object Detection From an Egocentric Perspective
Boyang Deng, Charles R Qi, Mahyar Najibi, Thomas Funkhouser, Yin Zhou, Dragomir Anguelov
3D object detection is a key module for safety-critical robotics applications such as autonomous driving. For these applications, we care most about how the detections affect the ego-agent’s behavior and safety (the egocentric perspective). Intuitively, we seek more accurate descriptions of object geometry when it’s more likely to interfere with the ego-agent’s motion trajectory. However, current detection metrics, based on box Intersection-over-Union (IoU), are object-centric and aren’t designed to capture the spatio-temporal relationship between objects and the ego-agent. To address this issue, we propose a new egocentric measure to evaluate 3D object detection, namely Support Distance Error (SDE). Our analysis based on SDE reveals that the egocentric detection quality is bounded by the coarse geometry of the bounding boxes. Given the insight that SDE would benefit from more accurate geometry descriptions, we propose to represent objects as amodal contours, specifically amodal star-shaped polygons, and devise a simple model, StarPoly, to predict such contours. Our experiments on the large-scale Waymo Open Dataset show that SDE better reflects the impact of detection quality on the ego-agent’s safety compared to IoU; and the estimated contours from StarPoly consistently improve the egocentric detection quality over recent 3D object detectors.
****
Optimizing Information-theoretical Generalization Bound via Anisotropic Noise of

 SGLD
Bohan Wang, Huishuai Zhang, Jieyu Zhang, Qi Meng, Wei Chen, Tie-Yan Liu
Recently, the information-theoretical framework has been proven to be able to obtain non-vacuous generalization bounds for large models trained by Stochastic Gradient Langevin Dynamics (SGLD) with isotropic noise. In this paper, we optimize the information-theoretical generalization bound by manipulating the noise structure in SGLD. We prove that with constraint to guarantee low empirical risk, the optimal noise covariance is the square root of the expected gradient covariance if both the prior and the posterior are jointly optimized. This validates that the optimal noise is quite close to the empirical gradient covariance. Technically, we develop a new information-theoretical bound that enables such an optimization analysis. We then apply matrix analysis to derive the form of optimal noise covariance. Presented constraint and results are validated by the empirical observations.
****
Addressing Algorithmic Disparity and Performance Inconsistency in Federated Learning
Sen Cui, Weishen Pan, Jian Liang, Changshui Zhang, Fei Wang
Federated learning (FL) has gain growing interests for its capability of learning from distributed data sources collectively without the need of accessing the raw data samples across different sources. So far FL research has mostly focused on improving the performance, how the algorithmic disparity will be impacted for the model learned from FL and the impact of algorithmic disparity on the utility inconsistency are largely unexplored. In this paper, we propose an FL framework to jointly consider performance consistency and algorithmic fairness across different local clients (data sources). We derive our framework from a constrained multi-objective optimization perspective, in which we learn a model satisfying fairness constraints on all clients with consistent performance. Specifically, we treat the algorithm prediction loss at each local client as an objective and maximize the worst-performing client with fairness constraints through optimizing a surrogate maximum function with all objectives involved. A gradient-based procedure is employed to achieve the Pareto optimality of this optimization problem. Theoretical analysis is provided to prove that our method can converge to a Pareto solution that achieves the min-max performance with fairness constraints on all clients. Comprehensive experiments on synthetic and real-world datasets demonstrate the superiority that our approach over baselines and its effectiveness in achieving both fairness and consistency across all local clients.
****
A Mathematical Framework for Quantifying Transferability in Multi-source Transfer Learning
Xinyi Tong, Xiangxiang Xu, Shao-Lun Huang, Lizhong Zheng
Current transfer learning algorithm designs mainly focus on the similarities between source and target tasks, while the impacts of the sample sizes of these tasks are often not sufficiently addressed. This paper proposes a mathematical framework for quantifying the transferability in multi-source transfer learning problems, with both the task similarities and the sample complexity of learning models taken into account. In particular, we consider the setup where the models learned from different tasks are linearly combined for learning the target task, and use the optimal combining coefficients to measure the transferability. Then, we demonstrate the analytical expression of this transferability measure, characterized by the sample sizes, model complexity, and the similarities between source and target tasks, which provides fundamental insights of the knowledge transferring mechanism and the guidance for algorithm designs. Furthermore, we apply our analyses for practical learning tasks, and establish a quantifiable transferability measure by exploiting a parameterized model. In addition, we develop an alternating iterative algorithm to implement our theoretical results for training deep neural networks in multi-source transfer learning tasks. Finally, experiments on image classification tasks show that our approach outperforms existing transfer learning algorithms in multi-source and few-shot scenarios.
****

Morié Attack (MA): A New Potential Risk of Screen Photos
Dantong Niu, Ruohao Guo, Yisen Wang
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Fast Bayesian Inference for Gaussian Cox Processes via Path Integral Formulation
Hideaki Kim
Gaussian Cox processes are widely-used point process models that use a Gaussian process to describe the Bayesian a priori uncertainty present in latent intensity functions. In this paper, we propose a novel Bayesian inference scheme for Gaussian Cox processes by exploiting a conceptually-intuitive {¥it path integral} formulation. The proposed scheme does not rely on domain discretization, scales linearly with the number of observed events, has a lower complexity than the state-of-the-art variational Bayesian schemes with respect to the number of inducing points, and is applicable to a wide range of Gaussian Cox processes with various types of link functions. Our scheme is especially beneficial under the multi-dimensional input setting, where the number of inducing points tends to be large. We evaluate our scheme on synthetic and real-world data, and show that it achieves comparable predictive accuracy while being tens of times faster than reference methods.
****
Lattice partition recovery with dyadic CART
OSCAR HERNAN MADRID PADILLA, Yi Yu, Alessandro Rinaldo
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 Deep Reinforcement Learning through Adversarial Loss
Tuomas Oikarinen, Wang Zhang, Alexandre Megretski, Luca Daniel, Tsui-Wei Weng
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Provable Model-based Nonlinear Bandit and Reinforcement Learning: Shelve Optimism, Embrace Virtual Curvature
Kefan Dong, Jiaqi Yang, Tengyu Ma
This paper studies model-based bandit and reinforcement learning (RL) with nonlinear function approximations. We propose to study convergence to approximate local maxima because we show that global convergence is statistically intractable even for one-layer neural net bandit with a deterministic reward. For both nonlinear bandit and RL, the paper presents a model-based algorithm, Virtual Ascent with Online Model Learner (ViOlin), which provably converges to a local maximum with sample complexity that only depends on the sequential Rademacher complexity of the model class. Our results imply novel global or local regret bounds on several concrete settings such as linear bandit with finite or sparse model class, and two-layer neural net bandit. A key algorithmic insight is that optimism may lead to over-exploration even for two-layer neural net model class. On the other hand, for convergence to local maxima, it suffices to maximize the virtual return if the model can also reasonably predict the gradient and Hessian of the real return.
****
You Only Look at One Sequence: Rethinking Transformer in Vision through Object Detection
Yuxin Fang, Bencheng Liao, Xinggang Wang, Jiemin Fang, Jiyang Qi, Rui Wu, Jianwei Niu, Wenyu 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.
****
Learning to delegate for large-scale vehicle routing
Sirui Li, Zhongxia Yan, Cathy 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.
****
Effective Meta-Regularization by Kernelized Proximal Regularization
Weisen Jiang, James Kwok, Yu Zhang
We study the problem of meta-learning, which has proved to be advantageous to accelerate learning new tasks with a few samples. The recent approaches based on deep kernels achieve the state-of-the-art performance. However, the regularizers in their base learners are not learnable. In this paper, we propose an algorithm called MetaProx to learn a proximal regularizer for the base learner. We theoretically establish the convergence of MetaProx. Experimental results confirm the advantage of the proposed algorithm.
****
Towards Context-Agnostic Learning Using Synthetic Data
Charles Jin, Martin Rinard
We propose a novel setting for learning, where the input domain is the image of a map defined on the product of two sets, one of which completely determines the labels. We derive a new risk bound for this setting that decomposes into a bias and an error term, and exhibits a surprisingly weak dependence on the true labels. Inspired by these results, we present an algorithm aimed at minimizing the bias term by exploiting the ability to sample from each set independently. We apply our setting to visual classification tasks, where our approach enables us to train classifiers on datasets that consist entirely of a single synthetic example of each class. On several standard benchmarks for real-world image classification, we achieve robust performance in the context-agnostic setting, with good generalization to real world domains, whereas training directly on real world data without our techniques yields classifiers that are brittle to perturbations of the background.
****
Minimax Optimal Quantile and Semi-Adversarial Regret via Root-Logarithmic Regularizers
Jeffrey Negrea, Blair Bilodeau, Nicolò Campolongo, Francesco Orabona, Dan Roy
Quantile (and, more generally, KL) regret bounds, such as those achieved by NormalHedge (Chaudhuri, Freund, and Hsu 2009) and its variants, relax the goal of competing against the best individual expert to only competing against a majority of experts on adversarial data. More recently, the semi-adversarial paradigm (Bilodeau, Negrea, and Roy 2020) provides an alternative relaxation of adversarial online learning by considering data that may be neither fully adversarial nor stochastic (I.I.D.). We achieve the minimax optimal regret in both paradigms using FTRL with separate, novel, root-logarithmic regularizers, both of which can be interpreted as yielding variants of NormalHedge. We extend existing KL regret upper bounds, which hold uniformly over target distributions, to possibly uncountable expert classes with arbitrary priors; provide the first full-information lower bounds for quantile regret on finite expert classes (which are tight); and provide an adaptively minimax optimal algorithm for the semi-adversarial paradigm that adapts to the true, unknown constraint faster, leading to uniformly improved regret bounds over existing methods.
****
Gradient-Free Adversarial Training Against Image Corruption for Learning-based Steering
Yu Shen, Laura Zheng, Manli Shu, Weizi Li, Tom Goldstein, Ming Lin
We introduce a simple yet effective framework for improving the robustness of le

arning algorithms against image corruptions for autonomous driving. These corrup tions can occur due to both internal (e.g., sensor noises and hardware abnormali ties) and external factors (e.g., lighting, weather, visibility, and other envir onmental effects). Using sensitivity analysis with FID-based parameterization, w e propose a novel algorithm exploiting basis perturbations to improve the overal l performance of autonomous steering and other image processing tasks, such as c lassification and detection, for self-driving cars. Our model not only improves the performance on the original dataset, but also achieves significant performan ce improvement on datasets with multiple and unseen perturbations, up to 87% and 77%, respectively. A comparison between our approach and other SOTA techniques confirms the effectiveness of our technique in improving the robustness of neura l network training for learning-based steering and other image processing tasks.
****
Deep Proxy Causal Learning and its Application to Confounded Bandit Policy Evalu ation
Liyuan Xu, Heishiro Kanagawa, Arthur Gretton
Proxy causal learning (PCL) is a method for estimating the causal effect of trea tments on outcomes in the presence of unobserved confounding, using proxies (str uctured side information) for the confounder. This is achieved via two-stage reg ression: in the first stage, we model relations among the treatment and proxies; in the second stage, we use this model to learn the effect of treatment on the outcome, given the context provided by the proxies. PCL guarantees recovery of the true causal effect, subject to identifiability conditions. We propose a nove l method for PCL, the deep feature proxy variable method (DFPV), to address the case where the proxies, treatments, and outcomes are high-dimensional and have n onlinear complex relationships, as represented by deep neural network features. We show that DFPV outperforms recent state-of-the-art PCL methods on challenging synthetic benchmarks, including settings involving high dimensional image data. Furthermore, we show that PCL can be applied to off-policy evaluation for the c onfounded bandit problem, in which DFPV also exhibits competitive performance.
****
Certifying Robustness to Programmable Data Bias in Decision Trees
Anna Meyer, Aws Albarghouthi, Loris D'Antoni
Datasets can be biased due to societal inequities, human biases, under-represent ation of minorities, etc. Our goal is to certify that models produced by a learn ing algorithm are pointwise-robust to dataset biases. This is a challenging prob lem: it entails learning models for a large, or even infinite, number of dataset s, ensuring that they all produce the same prediction. We focus on decision-tree learning due to the interpretable nature of the models. Our approach allows pro grammatically specifying \emph{bias models} across a variety of dimensions (e.g. , label-flipping or missing data), composing types of bias, and targeting bias t owards a specific group. To certify robustness, we use a novel symbolic techniqu e to evaluate a decision-tree learner on a large, or infinite, number of dataset s, certifying that each and every dataset produces the same prediction for a spe cific test point. We evaluate our approach on datasets that are commonly used in the fairness literature, and demonstrate our approach's viability on a range of bias models.
****
TöRF: Time-of-Flight Radiance Fields for Dynamic Scene View Synthesis
Benjamin Attal, Eliot Laidlaw, Aaron Gokaslan, Changil Kim, Christian Richardt, James Tompkin, Matthew O'Toole
Neural networks can represent and accurately reconstruct radiance fields for sta tic 3D scenes (e.g., NeRF). Several works extend these to dynamic scenes capture d with monocular video, with promising performance. However, the monocular setti ng is known to be an under-constrained problem, and so methods rely on data-driv en priors for reconstructing dynamic content. We replace these priors with measu rements from a time-of-flight (ToF) camera, and introduce a neural representatio n based on an image formation model for continuous-wave ToF cameras. Instead of working with processed depth maps, we model the raw ToF sensor measurements to i mprove reconstruction quality and avoid issues with low reflectance regions, mul

ti-path interference, and a sensor's limited unambiguous depth range. We show th
at this approach improves robustness of dynamic scene reconstruction to erroneou
s calibration and large motions, and discuss the benefits and limitations of int
egrating RGB+ToF sensors now available on modern smartphones.
****

Sequence-to-Sequence Learning with Latent Neural Grammars

Yoon Kim

Sequence-to-sequence learning with neural networks has become the de facto stand
ard for sequence modeling. This approach typically models the local distribution
 over the next element with a powerful neural network that can condition on arbi
trary context. While flexible and performant, these models often require large d
atasets for training and can fail spectacularly on benchmarks designed to test f
or compositional generalization. This work explores an alternative, hierarchical
 approach to sequence-to-sequence learning with synchronous grammars, where each
 node in the target tree is transduced by a subset of nodes in the source tree.
The source and target trees are treated as fully latent and marginalized out dur
ing training. We develop a neural parameterization of the grammar which enables
parameter sharing over combinatorial structures without the need for manual feat
ure engineering. We apply this latent neural grammar to various domains---a diag
nostic language navigation task designed to test for compositional generalizatio
n (SCAN), style transfer, and small-scale machine translation---and find that it
 performs respectably compared to standard baselines.
****

Exploration-Exploitation in Multi-Agent Competition: Convergence with Bounded Ra
tionality

Stefanos Leonardos, Georgios Piliouras, Kelly Spendlove

The interplay between exploration and exploitation in competitive multi-agent le
arning is still far from being well understood. Motivated by this, we study smoo
th Q-learning, a prototypical learning model that explicitly captures the balanc
e between game rewards and exploration costs. We show that Q-learning always con
verges to the unique quantal-response equilibrium (QRE), the standard solution c
oncept for games under bounded rationality, in weighted zero-sum polymatrix game
s with heterogeneous learning agents using positive exploration rates. Complemen
ting recent results about convergence in weighted potential games [16,34], we sh
ow that fast convergence of Q-learning in competitive settings obtains regardles
s of the number of agents and without any need for parameter fine-tuning. As sho
wcased by our experiments in network zero-sum games, these theoretical results p
rovide the necessary guarantees for an algorithmic approach to the currently ope
n problem of equilibrium selection in competitive multi-agent settings.
****

Low-Rank Extragradient Method for Nonsmooth and Low-Rank Matrix Optimization Pro
blems

Atara Kaplan, Dan Garber

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

Which Mutual-Information Representation Learning Objectives are Sufficient for C
ontrol?

Kate Rakelly, Abhishek Gupta, Carlos Florensa, Sergey Levine

Mutual information (MI) maximization provides an appealing formalism for learnin
g representations of data. In the context of reinforcement learning (RL), such r
epresentations can accelerate learning by discarding irrelevant and redundant in
formation, while retaining the information necessary for control. Much prior wor
k on these methods has addressed the practical difficulties of estimating MI fro
m samples of high-dimensional observations, while comparatively less is understo
od about which MI objectives yield representations that are sufficient for RL fr
om a theoretical perspective. In this paper, we formalize the sufficiency of a s
tate representation for learning and representing the optimal policy, and study

several popular MI based objectives through this lens. Surprisingly, we find that two of these objectives can yield insufficient representations given mild and common assumptions on the structure of the MDP. We corroborate our theoretical results with empirical experiments on a simulated game environment with visual observations.
****

A Geometric Perspective towards Neural Calibration via Sensitivity Decomposition
Junjiao Tian, Dylan Yung, Yen-Chang Hsu, Zsolt Kira
It is well known that vision classification models suffer from poor calibration in the face of data distribution shifts. In this paper, we take a geometric approach to this problem. We propose Geometric Sensitivity Decomposition (GSD) which decomposes the norm of a sample feature embedding and the angular similarity to a target classifier into an instance-dependent and an instance-independent com-ponent. The instance-dependent component captures the sensitive information about changes in the input while the instance-independent component represents the insensitive information serving solely to minimize the loss on the training dataset. Inspired by the decomposition, we analytically derive a simple extension to current softmax-linear models, which learns to disentangle the two components during training. On several common vision models, the disentangled model out-performs other calibration methods on standard calibration metrics in the face of out-of-distribution (OOD) data and corruption with significantly less complexity. Specifically, we surpass the current state of the art by 30.8% relative improvement on corrupted CIFAR100 in Expected Calibration Error.
****

Towards a Unified Information-Theoretic Framework for Generalization
Mahdi Haghifam, Gintare Karolina Dziugaite, Shay Moran, Dan Roy
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 decision-making under misspecified priors with applications to meta-learning
Max Simchowitz, Christopher Tosh, Akshay Krishnamurthy, Daniel J. Hsu, Thodoris Lykouris, Miro Dudik, Robert E. Schapire
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 Trees for Learning on Graphs
Rajat Talak, Siyi Hu, Lisa Peng, Luca Carlone
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

Enabling Fast Differentially Private SGD via Just-in-Time Compilation and Vectorization
Pranav Subramani, Nicholas Vadivelu, Gautam Kamath
A common pain point in differentially private machine learning is the significant runtime overhead incurred when executing Differentially Private Stochastic Gradient Descent (DPSGD), which may be as large as two orders of magnitude. We thoroughly demonstrate that by exploiting powerful language primitives, including vectorization, just-in-time compilation, and static graph optimization, one can dramatically reduce these overheads, in many cases nearly matching the best non-private running times. These gains are realized in two frameworks: one is JAX, which provides rich support for these primitives through the XLA compiler. We also rebuild core parts of TensorFlow Privacy, integrating more effective vectorization as well as XLA compilation, granting significant memory and runtime improveme

nts over previous release versions. Our proposed approaches allow us to achieve up to 50x speedups compared to the best alternatives. Our code is available at https://github.com/TheSalon/fast-dpsgd.
****
The effectiveness of feature attribution methods and its correlation with automatic evaluation scores
Giang Nguyen, Daeyoung Kim, Anh Nguyen
Explaining the decisions of an Artificial Intelligence (AI) model is increasingly critical in many real-world, high-stake applications.Hundreds of papers have either proposed new feature attribution methods, discussed or harnessed these tools in their work.However, despite humans being the target end-users, most attribution methods were only evaluated on proxy automatic-evaluation metrics (Zhang et al. 2018; Zhou et al. 2016; Petsiuk et al. 2018). In this paper, we conduct the first user study to measure attribution map effectiveness in assisting humans in ImageNet classification and Stanford Dogs fine-grained classification, and when an image is natural or adversarial (i.e., contains adversarial perturbations). Overall, feature attribution is surprisingly not more effective than showing humans nearest training-set examples. On a harder task of fine-grained dog categorization, presenting attribution maps to humans does not help, but instead hurts the performance of human-AI teams compared to AI alone. Importantly, we found automatic attribution-map evaluation measures to correlate poorly with the actual human-AI team performance. Our findings encourage the community to rigorously test their methods on the downstream human-in-the-loop applications and to rethink the existing evaluation metrics.
****
Coordinated Proximal Policy Optimization
Zifan Wu, Chao Yu, Deheng Ye, Junge Zhang, haiyin piao, Hankz Hankui Zhuo
We present Coordinated Proximal Policy Optimization (CoPPO), an algorithm that extends the original Proximal Policy Optimization (PPO) to the multi-agent setting. The key idea lies in the coordinated adaptation of step size during the policy update process among multiple agents. We prove the monotonicity of policy improvement when optimizing a theoretically-grounded joint objective, and derive a simplified optimization objective based on a set of approximations. We then interpret that such an objective in CoPPO can achieve dynamic credit assignment among agents, thereby alleviating the high variance issue during the concurrent update of agent policies. Finally, we demonstrate that CoPPO outperforms several strong baselines and is competitive with the latest multi-agent PPO method (i.e. MAPPO) under typical multi-agent settings, including cooperative matrix games and the StarCraft II micromanagement tasks.
****
Unbiased Classification through Bias-Contrastive and Bias-Balanced Learning
Youngkyu Hong, Eunho Yang
Datasets for training machine learning models tend to be biased unless the data is collected with complete care. In such a biased dataset, models are susceptible to making predictions based on the biased features of the data. The biased model fails to generalize to the case where correlations between biases and targets are shifted. To mitigate this, we propose Bias-Contrastive (BiasCon) loss based on the contrastive learning framework, which effectively leverages the knowledge of bias labels. We further suggest Bias-Balanced (BiasBal) regression which trains the classification model toward the data distribution with balanced target-bias correlation. Furthermore, we propose Soft Bias-Contrastive (SoftCon) loss which handles the dataset without bias labels by softening the pair assignment of the BiasCon loss based on the distance in the feature space of the bias-capturing model. Our experiments show that our proposed methods significantly improve previous debiasing methods in various realistic datasets.
****
Learning from Inside: Self-driven Siamese Sampling and Reasoning for Video Question Answering
Weijiang Yu, Haoteng Zheng, Mengfei Li, Lei Ji, Lijun Wu, Nong Xiao, Nan Duan
Recent advances in the video question answering (i.e., VideoQA) task have achiev

ed strong success by following the paradigm of fine-tuning each clip-text pair independently on the pretrained transformer-based model via supervised learning. Intuitively, multiple samples (i.e., clips) should be interdependent to capture similar visual and key semantic information in the same video. To consider the interdependent knowledge between contextual clips into the network inference, we propose a Siamese Sampling and Reasoning (SiaSamRea) approach, which consists of a siamese sampling mechanism to generate sparse and similar clips (i.e., siamese clips) from the same video, and a novel reasoning strategy for integrating the interdependent knowledge between contextual clips into the network. The reasoning strategy contains two modules: (1) siamese knowledge generation to learn the inter-relationship among clips; (2) siamese knowledge reasoning to produce the refined soft label by propagating the weights of inter-relationship to the predicted candidates of all clips. Finally, our SiaSamRea can endow the current multimodal reasoning paradigm with the ability of learning from inside via the guidance of soft labels. Extensive experiments demonstrate our SiaSamRea achieves state-of-the-art performance on five VideoQA benchmarks, e.g., a significant +2.1% gain on MSRVTT-QA, +2.9% on MSVD-QA, +1.0% on ActivityNet-QA, +1.8% on How2QA and +4.3% (action) on TGIF-QA.
****

Identification and Estimation of Joint Probabilities of Potential Outcomes in Observational Studies with Covariate Information

Ryusei Shingaki, manabu kuroki

The joint probabilities of potential outcomes are fundamental components of causal inference in the sense that (i) if they are identifiable, then the causal risk is also identifiable, but not vise versa (Pearl, 2009; Tian and Pearl, 2000) and (ii) they enable us to evaluate the probabilistic aspects of necessity'',sufficiency'', and ``necessity and sufficiency'', which are important concepts of successful explanation (Watson, et al., 2020). However, because they are not identifiable without any assumptions, various assumptions have been utilized to evaluate the joint probabilities of potential outcomes, e.g., the assumption of monotonicity (Pearl, 2009; Tian and Pearl, 2000), the independence between potential outcomes (Robins and Richardson, 2011), the condition of gain equality (Li and Pearl, 2019), and the specific functional relationships between cause and effect (Pearl, 2009). Unlike existing identification conditions, in order to evaluate the joint probabilities of potential outcomes without such assumptions, this paper proposes two types of novel identification conditions using covariate information. In addition, when the joint probabilities of potential outcomes are identifiable through the proposed conditions, the estimation problem of the joint probabilities of potential outcomes reduces to that of singular models and thus they can not be evaluated by standard statistical estimation methods. To solve the problem, this paper proposes a new statistical estimation method based on the augmented Lagrangian method and shows the asymptotic normality of the proposed estimators. Given space constraints, the proofs, the details on the statistical estimation method, some numerical experiments, and the case study are provided in the supplementary material.
****

Online false discovery rate control for anomaly detection in time series

Quentin Rebjock, Baris Kurt, Tim Januschowski, Laurent Callot

This article proposes novel rules for false discovery rate control (FDRC) geared towards online anomaly detection in time series. Online FDRC rules allow to control the properties of a sequence of statistical tests. In the context of anomaly detection, the null hypothesis is that an observation is normal and the alternative is that it is anomalous. FDRC rules allow users to target a lower bound on precision in unsupervised settings. The methods proposed in this article overcome short-comings of previous FDRC rules in the context of anomaly detection, in particular ensuring that power remains high even when the alternative is exceedingly rare (typical in anomaly detection) and the test statistics are serially dependent (typical in time series). We show the soundness of these rules in both theory and experiments.
****

Pragmatic Image Compression for Human-in-the-Loop Decision-Making
Sid Reddy, Anca Dragan, Sergey Levine
Standard lossy image compression algorithms aim to preserve an image's appearance, while minimizing the number of bits needed to transmit it. However, the amount of information actually needed by the user for downstream tasks -- e.g., deciding which product to click on in a shopping website -- is likely much lower. To achieve this lower bitrate, we would ideally only transmit the visual features that drive user behavior, while discarding details irrelevant to the user's decisions. We approach this problem by training a compression model through human-in-the-loop learning as the user performs tasks with the compressed images. The key insight is to train the model to produce a compressed image that induces the user to take the same action that they would have taken had they seen the original image. To approximate the loss function for this model, we train a discriminator that tries to distinguish whether a user's action was taken in response to the compressed image or the original. We evaluate our method through experiments with human participants on four tasks: reading handwritten digits, verifying photos of faces, browsing an online shopping catalogue, and playing a car racing video game. The results show that our method learns to match the user's actions with and without compression at lower bitrates than baseline methods, and adapts the compression model to the user's behavior: it preserves the digit number and randomizes handwriting style in the digit reading task, preserves hats and eyeglasses while randomizing faces in the photo verification task, preserves the perceived price of an item while randomizing its color and background in the online shopping task, and preserves upcoming bends in the road in the car racing game.
****
Generalized Linear Bandits with Local Differential Privacy
Yuxuan Han, Zhipeng Liang, Yang Wang, Jiheng 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.
****
On the Algorithmic Stability of Adversarial Training
Yue Xing, Qifan Song, Guang Cheng
The adversarial training is a popular tool to remedy the vulnerability of deep learning models against adversarial attacks, and there is rich theoretical literature on the training loss of adversarial training algorithms. In contrast, this paper studies the algorithmic stability of a generic adversarial training algorithm, which can further help to establish an upper bound for generalization error. By figuring out the stability upper bound and lower bound, we argue that the non-differentiability issue of adversarial training causes worse algorithmic stability than their natural counterparts. To tackle this problem, we consider a noise injection method. While the non-differentiability problem seriously affects the stability of adversarial training, injecting noise enables the training trajectory to avoid the occurrence of non-differentiability with dominating probability, hence enhancing the stability performance of adversarial training. Our analysis also studies the relation between the algorithm stability and numerical approximation error of adversarial attacks.
****
Width-based Lookaheads with Learnt Base Policies and Heuristics Over the Atari-2600 Benchmark
Stefan O'Toole, Nir Lipovetzky, Miquel Ramirez, Adrian Pearce
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Characterizing possible failure modes in physics-informed neural networks
Aditi Krishnapriyan, Amir Gholami, Shandian Zhe, Robert Kirby, Michael W. Mahoney

Recent work in scientific machine learning has developed so-called physics-infor
med neural network (PINN) models. The typical approach is to incorporate physica
l domain knowledge as soft constraints on an empirical loss function and use exi
sting machine learning methodologies to train the model. We demonstrate that, wh
ile existing PINN methodologies can learn good models for relatively trivial pro
blems, they can easily fail to learn relevant physical phenomena for even slight
ly more complex problems. In particular, we analyze several distinct situations
of widespread physical interest, including learning differential equations with
convection, reaction, and diffusion operators. We provide evidence that the soft
 regularization in PINNs, which involves PDE-based differential operators, can i
ntroduce a number of subtle problems, including making the problem more ill-cond
itioned. Importantly, we show that these possible failure modes are not due to t
he lack of expressivity in the NN architecture, but that the PINN's setup makes
the loss landscape very hard to optimize. We then describe two promising solutio
ns to address these failure modes. The first approach is to use curriculum regul
arization, where the PINN's loss term starts from a simple PDE regularization, a
nd becomes progressively more complex as the NN gets trained. The second approac
h is to pose the problem as a sequence-to-sequence learning task, rather than le
arning to predict the entire space-time at once. Extensive testing shows that we
 can achieve up to 1-2 orders of magnitude lower error with these methods as com
pared to regular PINN training.
****
Artistic Style Transfer with Internal-external Learning and Contrastive Learning
Haibo Chen, lei zhao, Zhizhong Wang, Huiming Zhang, Zhiwen Zuo, Ailin Li, Wei Xi
ng, Dongming Lu
Although existing artistic style transfer methods have achieved significant impr
ovement with deep neural networks, they still suffer from artifacts such as dish
armonious colors and repetitive patterns. Motivated by this, we propose an inter
nal-external style transfer method with two contrastive losses. Specifically, we
 utilize internal statistics of a single style image to determine the colors and
 texture patterns of the stylized image, and in the meantime, we leverage the ex
ternal information of the large-scale style dataset to learn the human-aware sty
le information, which makes the color distributions and texture patterns in the
stylized image more reasonable and harmonious. In addition, we argue that existi
ng style transfer methods only consider the content-to-stylization and style-to-
stylization relations, neglecting the stylization-to-stylization relations. To a
ddress this issue, we introduce two contrastive losses, which pull the multiple
stylization embeddings closer to each other when they share the same content or
style, but push far away otherwise. We conduct extensive experiments, showing th
at our proposed method can not only produce visually more harmonious and satisfy
ing artistic images, but also promote the stability and consistency of rendered
video clips.
****
Fast Abductive Learning by Similarity-based Consistency Optimization
Yu-Xuan Huang, Wang-Zhou Dai, Le-Wen Cai, Stephen H Muggleton, Yuan Jiang
To utilize the raw inputs and symbolic knowledge simultaneously, some recent neu
ro-symbolic learning methods use abduction, i.e., abductive reasoning, to integr
ate sub-symbolic perception and logical inference. While the perception model, e
.g., a neural network, outputs some facts that are inconsistent with the symboli
c background knowledge base, abduction can help revise the incorrect perceived f
acts by minimizing the inconsistency between them and the background knowledge.
However, to enable effective abduction, previous approaches need an initialized
perception model that discriminates the input raw instances. This limits the app
lication of these methods, as the discrimination ability is usually acquired fro
m a thorough pre-training when the raw inputs are difficult to classify. In this
 paper, we propose a novel abduction strategy, which leverages the similarity be
tween samples, rather than the output information by the perceptual neural netwo
rk, to guide the search in abduction. Based on this principle, we further presen
t ABductive Learning with Similarity (ABLSim) and apply it to some difficult neu
ro-symbolic learning tasks. Experiments show that the efficiency of ABLSim is si

gnificantly higher than the state-of-the-art neuro-symbolic methods, allowing it to achieve better performance with less labeled data and weaker domain knowledge.
****
To Beam Or Not To Beam: That is a Question of Cooperation for Language GANs
Thomas Scialom, Paul-Alexis Dray, Jacopo Staiano, Sylvain Lamprier, Benjamin Piwowarski
Due to the discrete nature of words, language GANs require to be optimized from rewards provided by discriminator networks, via reinforcement learning methods. This is a much harder setting than for continuous tasks, which enjoy gradient flows from discriminators to generators, usually leading to dramatic learning instabilities. However, we claim that this can be solved by making discriminator and generator networks cooperate to produce output sequences during training. These cooperative outputs, inherently built to obtain higher discrimination scores, not only provide denser rewards for training but also form a more compact artificial set for discriminator training, hence improving its accuracy and stability.In this paper, we show that our SelfGAN framework, built on this cooperative principle, outperforms Teacher Forcing and obtains state-of-the-art results on two challenging tasks, Summarization and Question Generation.
****
Shapley Residuals: Quantifying the limits of the Shapley value for explanations
Indra Kumar, Carlos Scheidegger, Suresh Venkatasubramanian, Sorelle Friedler
Popular feature importance techniques compute additive approximations to nonlinear models by first defining a cooperative game describing the value of different subsets of the model's features, then calculating the resulting game's Shapley values to attribute credit additively between the features. However, the specific modeling settings in which the Shapley values are a poor approximation for the true game have not been well-described. In this paper we utilize an interpretation of Shapley values as the result of an orthogonal projection between vector spaces to calculate a residual representing the kernel component of that projection. We provide an algorithm for computing these residuals, characterize different modeling settings based on the value of the residuals, and demonstrate that they capture information about model predictions that Shapley values cannot. Shapley residuals can thus act as a warning to practitioners against overestimating the degree to which Shapley-value-based explanations give them insight into a model.
****
The Elastic Lottery Ticket Hypothesis
Xiaohan Chen, Yu Cheng, Shuohang Wang, Zhe Gan, Jingjing Liu, Zhangyang Wang
Lottery Ticket Hypothesis (LTH) raises keen attention to identifying sparse trainable subnetworks, or winning tickets, which can be trained in isolation to achieve similar or even better performance compared to the full models. Despite many efforts being made, the most effective method to identify such winning tickets is still Iterative Magnitude-based Pruning (IMP), which is computationally expensive and has to be run thoroughly for every different network. A natural question that comes in is: can we “transform” the winning ticket found in one network to another with a different architecture, yielding a winning ticket for the latter at the beginning, without re-doing the expensive IMP? Answering this question is not only practically relevant for efficient “once-for-all” winning ticket ■nding, but also theoretically appealing for uncovering inherently scalable sparse patterns in networks. We conduct extensive experiments on CIFAR-10 and ImageNet, and propose a variety of strategies to tweak the winning tickets found from different networks of the same model family (e.g., ResNets). Based on these results, we articulate the Elastic Lottery Ticket Hypothesis (E-LTH): by mindfully replicating (or dropping) and re-ordering layers for one network, its corresponding winning ticket could be stretched (or squeezed) into a subnetwork for another deeper (or shallower) network from the same family, whose performance is nearly the same competitive as the latter’s winning ticket directly found by IMP. We have also extensively compared E-LTH with pruning-at-initialization and dynamic sparse training methods, as well as discussed the generalizability of E-LTH to diffe

rent model families, layer types, and across datasets. Code is available at https://github.com/VITA-Group/ElasticLTH.
****

Joint Inference for Neural Network Depth and Dropout Regularization
Kishan K C, Rui Li, MohammadMahdi Gilany
Dropout regularization methods prune a neural network's pre-determined backbone structure to avoid overfitting. However, a deep model still tends to be poorly calibrated with high confidence on incorrect predictions. We propose a unified Bayesian model selection method to jointly infer the most plausible network depth warranted by data, and perform dropout regularization simultaneously. In particular, to infer network depth we define a beta process over the number of hidden layers which allows it to go to infinity. Layer-wise activation probabilities induced by the beta process modulate neuron activation via binary vectors of a conjugate Bernoulli process. Experiments across domains show that by adapting network depth and dropout regularization to data, our method achieves superior performance comparing to state-of-the-art methods with well-calibrated uncertainty estimates. In continual learning, our method enables neural networks to dynamically evolve their depths to accommodate incrementally available data beyond their initial structures, and alleviate catastrophic forgetting.
****

Tractable Density Estimation on Learned Manifolds with Conformal Embedding Flows
Brendan Ross, Jesse Cresswell
Normalizing flows are generative models that provide tractable density estimation via an invertible transformation from a simple base distribution to a complex target distribution. However, this technique cannot directly model data supported on an unknown low-dimensional manifold, a common occurrence in real-world domains such as image data. Recent attempts to remedy this limitation have introduced geometric complications that defeat a central benefit of normalizing flows: exact density estimation. We recover this benefit with Conformal Embedding Flows, a framework for designing flows that learn manifolds with tractable densities. We argue that composing a standard flow with a trainable conformal embedding is the most natural way to model manifold-supported data. To this end, we present a series of conformal building blocks and apply them in experiments with synthetic and real-world data to demonstrate that flows can model manifold-supported distributions without sacrificing tractable likelihoods.
****

The Limits of Optimal Pricing in the Dark
Quinlan Dawkins, Minbiao Han, Haifeng Xu
A ubiquitous learning problem in today’s digital market is, during repeated interactions between a seller and a buyer, how a seller can gradually learn optimal pricing decisions based on the buyer’s past purchase responses. A fundamental challenge of learning in such a strategic setup is that the buyer will naturally have incentives to manipulate his responses in order to induce more favorable learning outcomes for him. To understand the limits of the seller’s learning when facing such a strategic and possibly manipulative buyer, we study a natural yet powerful buyer manipulation strategy. That is, before the pricing game starts, the buyer simply commits to “imitate” a different value function by pretending to always react optimally according to this imitative value function. We fully characterize the optimal imitative value function that the buyer should imitate as well as the resultant seller revenue and buyer surplus under this optimal buyer manipulation. Our characterizations reveal many useful insights about what happens at equilibrium. For example, a seller with concave production cost will obtain essentially 0 revenue at equilibrium whereas the revenue for a seller with convex production cost is the Bregman divergence of her cost function between no production and certain production. Finally, and importantly, we show that a more powerful class of pricing schemes does not necessarily increase, in fact, may be harmful to, the seller’s revenue. Our results not only lead to an effective prescriptive way for buyers to manipulate learning algorithms but also shed lights on the limits of what a seller can really achieve when pricing in the dark.
****

No RL, No Simulation: Learning to Navigate without Navigating
Meera Hahn, Devendra Singh Chaplot, Shubham Tulsiani, Mustafa Mukadam, James M. Rehg, Abhinav Gupta
Most prior methods for learning navigation policies require access to simulation environments, as they need online policy interaction and rely on ground-truth maps for rewards. However, building simulators is expensive (requires manual effort for each and every scene) and creates challenges in transferring learned policies to robotic platforms in the real-world, due to the sim-to-real domain gap. In this paper, we pose a simple question: Do we really need active interaction, ground-truth maps or even reinforcement-learning (RL) in order to solve the image-goal navigation task? We propose a self-supervised approach to learn to navigate from only passive videos of roaming. Our approach, No RL, No Simulator (NRNS), is simple and scalable, yet highly effective. NRNS outperforms RL-based formulations by a significant margin. We present NRNS as a strong baseline for any future image-based navigation tasks that use RL or Simulation.
****
Analogous to Evolutionary Algorithm: Designing a Unified Sequence Model
Jiangning Zhang, Chao Xu, Jian Li, Wenzhou Chen, Yabiao Wang, Ying Tai, Shuo Chen, Chengjie Wang, Feiyue Huang, Yong Liu
Inspired by biological evolution, we explain the rationality of Vision Transformer by analogy with the proven practical Evolutionary Algorithm (EA) and derive that both of them have consistent mathematical representation. Analogous to the dynamic local population in EA, we improve the existing transformer structure and propose a more efficient EAT model, and design task-related heads to deal with different tasks more flexibly. Moreover, we introduce the spatial-filling curve into the current vision transformer to sequence image data into a uniform sequential format. Thus we can design a unified EAT framework to address multi-modal tasks, separating the network architecture from the data format adaptation. Our approach achieves state-of-the-art results on the ImageNet classification task compared with recent vision transformer works while having smaller parameters and greater throughput. We further conduct multi-modal tasks to demonstrate the superiority of the unified EAT, \eg, Text-Based Image Retrieval, and our approach improves the rank-1 by +3.7 points over the baseline on the CSS dataset.
****
Improving Compositionality of Neural Networks by Decoding Representations to Inputs
Mike Wu, Noah Goodman, Stefano Ermon
In traditional software programs, it is easy to trace program logic from variables back to input, apply assertion statements to block erroneous behavior, and compose programs together. Although deep learning programs have demonstrated strong performance on novel applications, they sacrifice many of the functionalities of traditional software programs. With this as motivation, we take a modest first step towards improving deep learning programs by jointly training a generative model to constrain neural network activations to "decode" back to inputs. We call this design a Decodable Neural Network, or DecNN. Doing so enables a form of compositionality in neural networks, where one can recursively compose DecNN with itself to create an ensemble-like model with uncertainty. In our experiments, we demonstrate applications of this uncertainty to out-of-distribution detection, adversarial example detection, and calibration --- while matching standard neural networks in accuracy. We further explore this compositionality by combining DecNN with pretrained models, where we show promising results that neural networks can be regularized from using protected features.
****
The Hardness Analysis of Thompson Sampling for Combinatorial Semi-bandits with Greedy Oracle
Fang Kong, Yueran Yang, Wei Chen, Shuai 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.

****
Universal Semi-Supervised Learning
Zhuo Huang, Chao Xue, Bo Han, Jian Yang, Chen Gong
Universal Semi-Supervised Learning (UniSSL) aims to solve the open-set problem where both the class distribution (i.e., class set) and feature distribution (i.e., feature domain) are different between labeled dataset and unlabeled dataset. Such a problem seriously hinders the realistic landing of classical SSL. Different from the existing SSL methods targeting at the open-set problem that only study one certain scenario of class distribution mismatch and ignore the feature distribution mismatch, we consider a more general case where a mismatch exists in both class and feature distribution. In this case, we propose a ''Class-shAring data detection and Feature Adaptation'' (CAFA) framework which requires no prior knowledge of the class relationship between the labeled dataset and unlabeled dataset. Particularly, CAFA utilizes a novel scoring strategy to detect the data in the shared class set. Then, it conducts domain adaptation to fully exploit the value of the detected class-sharing data for better semi-supervised consistency training. Exhaustive experiments on several benchmark datasets show the effectiveness of our method in tackling open-set problems.
****
Improving Deep Learning Interpretability by Saliency Guided Training
Aya Abdelsalam Ismail, Hector Corrada Bravo, Soheil Feizi
Saliency methods have been widely used to highlight important input features in model predictions. Most existing methods use backpropagation on a modified gradient function to generate saliency maps. Thus, noisy gradients can result in unfaithful feature attributions. In this paper, we tackle this issue and introduce a {\it saliency guided training} procedure for neural networks to reduce noisy gradients used in predictions while retaining the predictive performance of the model. Our saliency guided training procedure iteratively masks features with small and potentially noisy gradients while maximizing the similarity of model outputs for both masked and unmasked inputs. We apply the saliency guided training procedure to various synthetic and real data sets from computer vision, natural language processing, and time series across diverse neural architectures, including Recurrent Neural Networks, Convolutional Networks, and Transformers. Through qualitative and quantitative evaluations, we show that saliency guided training procedure significantly improves model interpretability across various domains while preserving its predictive performance.
****
SurvITE: Learning Heterogeneous Treatment Effects from Time-to-Event Data
Alicia Curth, Changhee Lee, Mihaela van der Schaar
We study the problem of inferring heterogeneous treatment effects from time-to-event data. While both the related problems of (i) estimating treatment effects for binary or continuous outcomes and (ii) predicting survival outcomes have been well studied in the recent machine learning literature, their combination -- albeit of high practical relevance -- has received considerably less attention. With the ultimate goal of reliably estimating the effects of treatments on instantaneous risk and survival probabilities, we focus on the problem of learning (discrete-time) treatment-specific conditional hazard functions. We find that unique challenges arise in this context due to a variety of covariate shift issues that go beyond a mere combination of well-studied confounding and censoring biases. We theoretically analyse their effects by adapting recent generalization bounds from domain adaptation and treatment effect estimation to our setting and discuss implications for model design. We use the resulting insights to propose a novel deep learning method for treatment-specific hazard estimation based on balancing representations. We investigate performance across a range of experimental settings and empirically confirm that our method outperforms baselines by addressing covariate shifts from various sources.
****
Optimal Rates for Nonparametric Density Estimation under Communication Constraints
Jayadev Acharya, Clement Canonne, Aditya Vikram Singh, Himanshu Tyagi

We consider density estimation for Besov spaces when the estimator is restricted to use only a limited number of bits about each sample. We provide a noninteractive adaptive estimator which exploits the sparsity of wavelet bases, along with a simulate-and-infer technique from parametric estimation under communication constraints. We show that our estimator is nearly rate-optimal by deriving minmax lower bounds that hold even when interactive protocols are allowed. Interestingly, while our wavelet-based estimator is almost rate-optimal for Sobolev spaces as well, it is unclear whether the standard Fourier basis, which arise naturally for those spaces, can be used to achieve the same performance.
****
Rank Overspecified Robust Matrix Recovery: Subgradient Method and Exact Recovery
Lijun Ding, Liwei Jiang, Yudong Chen, Qing Qu, Zhihui Zhu
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Improving Computational Efficiency in Visual Reinforcement Learning via Stored Embeddings
Lili Chen, Kimin Lee, Aravind Srinivas, Pieter Abbeel
Recent advances in off-policy deep reinforcement learning (RL) have led to impressive success in complex tasks from visual observations. Experience replay improves sample-efficiency by reusing experiences from the past, and convolutional neural networks (CNNs) process high-dimensional inputs effectively. However, such techniques demand high memory and computational bandwidth. In this paper, we present Stored Embeddings for Efficient Reinforcement Learning (SEER), a simple modification of existing off-policy RL methods, to address these computational and memory requirements. To reduce the computational overhead of gradient updates in CNNs, we freeze the lower layers of CNN encoders early in training due to early convergence of their parameters. Additionally, we reduce memory requirements by storing the low-dimensional latent vectors for experience replay instead of high-dimensional images, enabling an adaptive increase in the replay buffer capacity, a useful technique in constrained-memory settings. In our experiments, we show that SEER does not degrade the performance of RL agents while significantly saving computation and memory across a diverse set of DeepMind Control environments and Atari games.
****
Learning Generalized Gumbel-max Causal Mechanisms
Guy Lorberbom, Daniel D. Johnson, Chris J. Maddison, Daniel Tarlow, Tamir Hazan
To perform counterfactual reasoning in Structural Causal Models (SCMs), one needs to know the causal mechanisms, which provide factorizations of conditional distributions into noise sources and deterministic functions mapping realizations of noise to samples. Unfortunately, the causal mechanism is not uniquely identified by data that can be gathered by observing and interacting with the world, so there remains the question of how to choose causal mechanisms. In recent work, Oberst & Sontag (2019) propose Gumbel-max SCMs, which use Gumbel-max reparameterizations as the causal mechanism due to an appealing counterfactual stability property. However, the justification requires appealing to intuition. In this work, we instead argue for choosing a causal mechanism that is best under a quantitative criteria such as minimizing variance when estimating counterfactual treatment effects. We propose a parameterized family of causal mechanisms that generalize Gumbel-max. We show that they can be trained to minimize counterfactual effect variance and other losses on a distribution of queries of interest, yielding lower variance estimates of counterfactual treatment effect than fixed alternatives, also generalizing to queries not seen at training time.
****
Bandit Learning with Delayed Impact of Actions
Wei Tang, Chien-Ju Ho, Yang 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.

****

A Stochastic Newton Algorithm for Distributed Convex Optimization

Brian Bullins, Kshitij Patel, Ohad Shamir, Nathan Srebro, Blake E. Woodworth

We propose and analyze a stochastic Newton algorithm for homogeneous distributed stochastic convex optimization, where each machine can calculate stochastic gradients of the same population objective, as well as stochastic Hessian-vector products (products of an independent unbiased estimator of the Hessian of the population objective with arbitrary vectors), with many such stochastic computations performed between rounds of communication. We show that our method can reduce the number, and frequency, of required communication rounds, compared to existing methods without hurting performance, by proving convergence guarantees for quasi-self-concordant objectives (e.g., logistic regression), alongside empirical evidence.

****

Are Transformers more robust than CNNs?

Yutong Bai, Jieru Mei, Alan L. Yuille, Cihang Xie

Transformer emerges as a powerful tool for visual recognition. In addition to demonstrating competitive performance on a broad range of visual benchmarks, recent works also argue that Transformers are much more robust than Convolutions Neural Networks (CNNs). Nonetheless, surprisingly, we find these conclusions are drawn from unfair experimental settings, where Transformers and CNNs are compared at different scales and are applied with distinct training frameworks. In this paper, we aim to provide the first fair & in-depth comparisons between Transformers and CNNs, focusing on robustness evaluations. With our unified training setup, we first challenge the previous belief that Transformers outshine CNNs when measuring adversarial robustness. More surprisingly, we find CNNs can easily be as robust as Transformers on defending against adversarial attacks, if they properly adopt Transformers' training recipes. While regarding generalization on out-of-distribution samples, we show pre-training on (external) large-scale datasets is not a fundamental request for enabling Transformers to achieve better performance than CNNs. Moreover, our ablations suggest such stronger generalization is largely benefited by the Transformer's self-attention-like architectures per se, rather than by other training setups. We hope this work can help the community better understand and benchmark the robustness of Transformers and CNNs. The code and models are publicly available at: https://github.com/ytongbai/ViTs-vs-CNNs.

****

Towards Sharper Generalization Bounds for Structured Prediction

Shaojie Li, Yong 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.

****

Automated Discovery of Adaptive Attacks on Adversarial Defenses

Chengyuan Yao, Pavol Bielik, Petar Tsankov, Martin Vechev

Reliable evaluation of adversarial defenses is a challenging task, currently limited to an expert who manually crafts attacks that exploit the defense’s inner workings, or to approaches based on ensemble of fixed attacks, none of which may be effective for the specific defense at hand. Our key observation is that adaptive attacks are composed from a set of reusable building blocks that can be formalized in a search space and used to automatically discover attacks for unknown defenses. We evaluated our approach on 24 adversarial defenses and show that it outperforms AutoAttack, the current state-of-the-art tool for reliable evaluation of adversarial defenses: our tool discovered significantly stronger attacks by producing 3.0%-50.8% additional adversarial examples for 10 models, while obtaining attacks with slightly stronger or similar strength for the remaining models.

****
PolarStream: Streaming Object Detection and Segmentation with Polar Pillars
Qi Chen, Sourabh Vora, Oscar Beijbom
Recent works recognized lidars as an inherently streaming data source and showed that the end-to-end latency of lidar perception models can be reduced significantly by operating on wedge-shaped point cloud sectors rather then the full point cloud. However, due to use of cartesian coordinate systems these methods represent the sectors as rectangular regions, wasting memory and compute. In this work we propose using a polar coordinate system and make two key improvements on this design. First, we increase the spatial context by using multi-scale padding from neighboring sectors: preceding sector from the current scan and/or the following sector from the past scan. Second, we improve the core polar convolutional architecture by introducing feature undistortion and range stratified convolutions. Experimental results on the nuScenes dataset show significant improvements over other streaming based methods. We also achieve comparable results to existing non-streaming methods but with lower latencies.
****
Representation Costs of Linear Neural Networks: Analysis and Design
Zhen Dai, Mina Karzand, Nathan Srebro
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Teaching via Best-Case Counterexamples in the Learning-with-Equivalence-Queries Paradigm
Akash Kumar, Yuxin Chen, Adish Singla
We study the sample complexity of teaching, termed as "teaching dimension" (TD) in the literature, for the learning-with-equivalence-queries (LwEQ) paradigm. More concretely, we consider a learner who asks equivalence queries (i.e., "is the queried hypothesis the target hypothesis?"), and a teacher responds either "yes" or "no" along with a counterexample to the queried hypothesis. This learning paradigm has been extensively studied when the learner receives worst-case or random counterexamples; in this paper, we consider the optimal teacher who picks best-case counterexamples to teach the target hypothesis within a hypothesis class. For this optimal teacher, we introduce LwEQ-TD, a notion of TD capturing the teaching complexity (i.e., the number of queries made) in this paradigm. We show that a significant reduction in queries can be achieved with best-case counterexamples, in contrast to worst-case or random counterexamples, for different hypothesis classes. Furthermore, we establish new connections of LwEQ-TD to the well-studied notions of TD in the learning-from-samples paradigm.
****
Distilling Meta Knowledge on Heterogeneous Graph for Illicit Drug Trafficker Detection on Social Media
Yiyue Qian, Yiming Zhang, Yanfang (Fa Ye, Chuxu Zhang
Driven by the considerable profits, the crime of drug trafficking (a.k.a. illicit drug trading) has co-evolved with modern technologies, e.g., social media such as Instagram has become a popular platform for marketing and selling illicit drugs. The activities of online drug trafficking are nimble and resilient, which call for novel techniques to effectively detect, disrupt, and dismantle illicit drug trades. In this paper, we propose a holistic framework named MetaHG to automatically detect illicit drug traffickers on social media (i.e., Instagram), by tackling the following two new challenges: (1) different from existing works which merely focus on analyzing post content, MetaHG is capable of jointly modeling multi-modal content and relational structured information on social media for illicit drug trafficker detection; (2) in addition, through the proposed meta-learning technique, MetaHG addresses the issue of requiring sufficient data for model training. More specifically, in our proposed MetaHG, we first build a heterogeneous graph (HG) to comprehensively characterize the complex ecosystem of drug trafficking on social media. Then, we employ a relation-based graph convolutiona

l neural network to learn node (i.e., user) representations over the built HG, i
n which we introduce graph structure refinement to compensate the sparse connect
ion among entities in the HG for more robust node representation learning. After
wards, we propose a meta-learning algorithm for model optimization. A self-super
vised module and a knowledge distillation module are further designed to exploit
 unlabeled data for improving the model. Extensive experiments based on the real
-world data collected from Instagram demonstrate that the proposed MetaHG outper
forms state-of-the-art methods.
****
Curriculum Disentangled Recommendation with Noisy Multi-feedback
Hong Chen, Yudong Chen, Xin Wang, Ruobing Xie, Rui Wang, Feng Xia, Wenwu Zhu
Learning disentangled representations for user intentions from multi-feedback (i
.e., positive and negative feedback) can enhance the accuracy and explainability
 of recommendation algorithms. However, learning such disentangled representatio
ns from multi-feedback data is challenging because i) multi-feedback is complex
: there exist complex relations among different types of feedback (e.g., click,
unclick, and dislike, etc) as well as various user intentions, and ii) multi-fee
dback is noisy: there exists noisy (useless) information both in features and la
bels, which may deteriorate the recommendation performance. Existing works on d
isentangled representation learning only focus on positive feedback, failing to
handle the complex relations and noise hidden in multi-feedback data. To solve t
his problem, in this work we propose a Curriculum Disentangled Recommendation (C
DR) model that is capable of efficiently learning disentangled representations f
rom complex and noisy multi-feedback for better recommendation. Concretely, we d
esign a co-filtering dynamic routing mechanism that simultaneously captures the
complex relations among different behavioral feedback and user intentions as wel
l as denoise the representations in the feature level. We then present an adjust
able self-evaluating curriculum that is able to evaluate sample difficulties for
 better model training and conduct denoising in the label level via disregarding
 useless information. Our extensive experiments on several real-world datasets d
emonstrate that the proposed CDR model can significantly outperform several stat
e-of-the-art methods in terms of recommendation accuracy.
****
Interpretable agent communication from scratch (with a generic visual processor
emerging on the side)
Roberto Dessi, Eugene Kharitonov, Baroni Marco
As deep networks begin to be deployed as autonomous agents, the issue of how the
y can communicate with each other becomes important. Here, we train two deep net
s from scratch to perform realistic referent identification through unsupervised
 emergent communication. We show that the largely interpretable emergent protoco
l allows the nets to successfully communicate even about object types they did n
ot see at training time. The visual representations induced as a by-product of o
ur training regime, moreover, show comparable quality, when re-used as generic v
isual features, to a recent self-supervised learning model. Our results provide
concrete evidence of the viability of (interpretable) emergent deep net communic
ation in a more realistic scenario than previously considered, as well as establ
ishing an intriguing link between this field and self-supervised visual learning
.
****
MAU: A Motion-Aware Unit for Video Prediction and Beyond
Zheng Chang, Xinfeng Zhang, Shanshe Wang, Siwei Ma, Yan Ye, Xiang Xinguang, Wen
Gao
Accurately predicting inter-frame motion information plays a key role in video p
rediction tasks. In this paper, we propose a Motion-Aware Unit (MAU) to capture
reliable inter-frame motion information by broadening the temporal receptive fie
ld of the predictive units. The MAU consists of two modules, the attention modul
e and the fusion module. The attention module aims to learn an attention map bas
ed on the correlations between the current spatial state and the historical spat
ial states. Based on the learned attention map, the historical temporal states a
re aggregated to an augmented motion information (AMI). In this way, the predict

ive unit can perceive more temporal dynamics from a wider receptive field. Then, the fusion module is utilized to further aggregate the augmented motion information (AMI) and current appearance information (current spatial state) to the final predicted frame. The computation load of MAU is relatively low and the proposed unit can be easily applied to other predictive models. Moreover, an information recalling scheme is employed into the encoders and decoders to help preserve the visual details of the predictions. We evaluate the MAU on both video prediction and early action recognition tasks. Experimental results show that the MAU outperforms the state-of-the-art methods on both tasks.
****
Successor Feature Landmarks for Long-Horizon Goal-Conditioned Reinforcement Learning
Christopher Hoang, Sungryull Sohn, Jongwook Choi, Wilka Carvalho, Honglak Lee
Operating in the real-world often requires agents to learn about a complex environment and apply this understanding to achieve a breadth of goals. This problem, known as goal-conditioned reinforcement learning (GCRL), becomes especially challenging for long-horizon goals. Current methods have tackled this problem by augmenting goal-conditioned policies with graph-based planning algorithms. However, they struggle to scale to large, high-dimensional state spaces and assume access to exploration mechanisms for efficiently collecting training data. In this work, we introduce Successor Feature Landmarks (SFL), a framework for exploring large, high-dimensional environments so as to obtain a policy that is proficient for any goal. SFL leverages the ability of successor features (SF) to capture transition dynamics, using it to drive exploration by estimating state-novelty and to enable high-level planning by abstracting the state-space as a non-parametric landmark-based graph. We further exploit SF to directly compute a goal-conditioned policy for inter-landmark traversal, which we use to execute plans to "frontier" landmarks at the edge of the explored state space. We show in our experiments on MiniGrid and ViZDoom that SFL enables efficient exploration of large, high-dimensional state spaces and outperforms state-of-the-art baselines on long-horizon GCRL tasks.
****
Streaming Belief Propagation for Community Detection
Yuchen Wu, Jakab Tardos, Mohammadhossein Bateni, André Linhares, Filipe Miguel Goncalves de Almeida, Andrea Montanari, Ashkan Norouzi-Fard
The community detection problem requires to cluster the nodes of a network into a small number of well-connected ‘communities’. There has been substantial recent progress in characterizing the fundamental statistical limits of community detection under simple stochastic block models. However, in real-world applications, the network structure is typically dynamic, with nodes that join over time. In this setting, we would like a detection algorithm to perform only a limited number of updates at each node arrival. While standard voting approaches satisfy this constraint, it is unclear whether they exploit the network information optimally. We introduce a simple model for networks growing over time which we refer to as streaming stochastic block model (StSBM). Within this model, we prove that voting algorithms have fundamental limitations. We also develop a streaming belief-propagation (STREAMBP) approach, for which we prove optimality in certain regimes. We validate our theoretical findings on synthetic and real data
****
The staircase property: How hierarchical structure can guide deep learning
Emmanuel Abbe, Enric Boix-Adsera, Matthew S Brennan, Guy Bresler, Dheeraj Nagaraj
This paper identifies a structural property of data distributions that enables deep neural networks to learn hierarchically. We define the ``staircase'' property for functions over the Boolean hypercube, which posits that high-order Fourier coefficients are reachable from lower-order Fourier coefficients along increasing chains. We prove that functions satisfying this property can be learned in polynomial time using layerwise stochastic coordinate descent on regular neural networks -- a class of network architectures and initializations that have homogeneity properties. Our analysis shows that for such staircase functions and neur

al networks, the gradient-based algorithm learns high-level features by greedily combining lower-level features along the depth of the network. We further back our theoretical results with experiments showing that staircase functions are learnable by more standard ResNet architectures with stochastic gradient descent. Both the theoretical and experimental results support the fact that the staircase property has a role to play in understanding the capabilities of gradient-based learning on regular networks, in contrast to general polynomial-size networks that can emulate any Statistical Query or PAC algorithm, as recently shown.
****

MagNet: A Neural Network for Directed Graphs
Xitong Zhang, Yixuan He, Nathan Brugnone, Michael Perlmutter, Matthew Hirn
The prevalence of graph-based data has spurred the rapid development of graph neural networks (GNNs) and related machine learning algorithms. Yet, despite the many datasets naturally modeled as directed graphs, including citation, website, and traffic networks, the vast majority of this research focuses on undirected graphs. In this paper, we propose MagNet, a GNN for directed graphs based on a complex Hermitian matrix known as the magnetic Laplacian. This matrix encodes undirected geometric structure in the magnitude of its entries and directional information in their phase. A charge parameter attunes spectral information to variation among directed cycles. We apply our network to a variety of directed graph node classification and link prediction tasks showing that MagNet performs well on all tasks and that its performance exceeds all other methods on a majority of such tasks. The underlying principles of MagNet are such that it can be adapted to other GNN architectures.
****

Hardware-adaptive Efficient Latency Prediction for NAS via Meta-Learning
Hayeon Lee, Sewoong Lee, Song Chong, Sung Ju Hwang
For deployment, neural architecture search should be hardware-aware, in order to satisfy the device-specific constraints (e.g., memory usage, latency and energy consumption) and enhance the model efficiency. Existing methods on hardware-aware NAS collect a large number of samples (e.g., accuracy and latency) from a target device, either builds a lookup table or a latency estimator. However, such approach is impractical in real-world scenarios as there exist numerous devices with different hardware specifications, and collecting samples from such a large number of devices will require prohibitive computational and monetary cost. To overcome such limitations, we propose Hardware-adaptive Efficient Latency Predictor (HELP), which formulates the device-specific latency estimation problem as a meta-learning problem, such that we can estimate the latency of a model's performance for a given task on an unseen device with a few samples. To this end, we introduce novel hardware embeddings to embed any devices considering them as black-box functions that output latencies, and meta-learn the hardware-adaptive latency predictor in a device-dependent manner, using the hardware embeddings. We validate the proposed HELP for its latency estimation performance on unseen platforms, on which it achieves high estimation performance with as few as 10 measurement samples, outperforming all relevant baselines. We also validate end-to-end NAS frameworks using HELP against ones without it, and show that it largely reduces the total time cost of the base NAS method, in latency-constrained settings.
****

Topological Relational Learning on Graphs
Yuzhou Chen, Baris Coskunuzer, Yulia Gel
Graph neural networks (GNNs) have emerged as a powerful tool for graph classification and representation learning. However, GNNs tend to suffer from over-smoothing problems and are vulnerable to graph perturbations. To address these challenges, we propose a novel topological neural framework of topological relational inference (TRI) which allows for integrating higher-order graph information to GNNs and for systematically learning a local graph structure. The key idea is to rewire the original graph by using the persistent homology of the small neighborhoods of the nodes and then to incorporate the extracted topological summaries as the side information into the local algorithm. As a result, the new framework enables us to harness both the conventional information on the graph structure an

d information on higher order topological properties of the graph. We derive the oretical properties on stability of the new local topological representation of the graph and discuss its implications on the graph algebraic connectivity. The experimental results on node classification tasks demonstrate that the new TRI-GNN outperforms all 14 state-of-the-art baselines on 6 out 7 graphs and exhibit higher robustness to perturbations, yielding up to 10\% better performance under noisy scenarios.
****

Learning Theory Can (Sometimes) Explain Generalisation in Graph Neural Networks
Pascal Esser, Leena Chennuru Vankadara, Debarghya Ghoshdastidar
In recent years, several results in the supervised learning setting suggested that classical statistical learning-theoretic measures, such as VC dimension, do not adequately explain the performance of deep learning models which prompted a slew of work in the infinite-width and iteration regimes. However, there is little theoretical explanation for the success of neural networks beyond the supervised setting. In this paper we argue that, under some distributional assumptions, classical learning-theoretic measures can sufficiently explain generalization for graph neural networks in the transductive setting. In particular, we provide a rigorous analysis of the performance of neural networks in the context of transductive inference, specifically by analysing the generalisation properties of graph convolutional networks for the problem of node classification. While VC-dimension does result in trivial generalisation error bounds in this setting as well, we show that transductive Rademacher complexity can explain the generalisation properties of graph convolutional networks for stochastic block models. We further use the generalisation error bounds based on transductive Rademacher complexity to demonstrate the role of graph convolutions and network architectures in achieving smaller generalisation error and provide insights into when the graph structure can help in learning. The findings of this paper could re-new the interest in studying generalisation in neural networks in terms of learning-theoretic measures, albeit in specific problems.
****

Federated Linear Contextual Bandits
Ruiquan Huang, Weiqiang Wu, Jing Yang, Cong Shen
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

Least Square Calibration for Peer Reviews
Sijun Tan, Jibang Wu, Xiaohui Bei, Haifeng Xu
Peer review systems such as conference paper review often suffer from the issue of miscalibration. Previous works on peer review calibration usually only use the ordinal information or assume simplistic reviewer scoring functions such as linear functions. In practice, applications like academic conferences often rely on manual methods, such as open discussions, to mitigate miscalibration. It remains an important question to develop algorithms that can   handle different types of miscalibrations based on available prior knowledge. In this paper, we propose a flexible framework, namely \emph{least square calibration} (LSC), for selecting top candidates from peer ratings. Our framework provably performs perfect calibration from noiseless linear scoring functions under mild assumptions, yet also provides competitive calibration results when the scoring function is from broader classes beyond linear functions and with arbitrary noise. On our synthetic dataset, we empirically demonstrate that our algorithm consistently outperforms the baseline which select top papers based on the highest average ratings.
****

Scaling Up Exact Neural Network Compression by ReLU Stability
Thiago Serra, Xin Yu, Abhinav Kumar, Srikumar Ramalingam
We can compress a rectifier network while exactly preserving its underlying functionality with respect to a given input domain if some of its neurons are stable. However, current approaches to determine the stability of neurons with Rectifi

ed Linear Unit (ReLU) activations require solving or finding a good approximation to multiple discrete optimization problems. In this work, we introduce an algorithm based on solving a single optimization problem to identify all stable neurons. Our approach is on median 183 times faster than the state-of-art method on CIFAR-10, which allows us to explore exact compression on deeper (5 x 100) and wider (2 x 800) networks within minutes. For classifiers trained under an amount of L1 regularization that does not worsen accuracy, we can remove up to 56% of the connections on the CIFAR-10 dataset. The code is available at the following link, https://github.com/yuxwind/ExactCompression .
****

Passive attention in artificial neural networks predicts human visual selectivity

Thomas Langlois, Haicheng Zhao, Erin Grant, Ishita Dasgupta, Tom Griffiths, Nori Jacoby

Developments in machine learning interpretability techniques over the past decade have provided new tools to observe the image regions that are most informative for classification and localization in artificial neural networks (ANNs). Are the same regions similarly informative to human observers? Using data from 79 new experiments and 7,810 participants, we show that passive attention techniques reveal a significant overlap with human visual selectivity estimates derived from 6 distinct behavioral tasks including visual discrimination, spatial localization, recognizability, free-viewing, cued-object search, and saliency search fixations. We find that input visualizations derived from relatively simple ANN architectures probed using guided backpropagation methods are the best predictors of a shared component in the joint variability of the human measures. We validate these correlational results with causal manipulations using recognition experiments. We show that images masked with ANN attention maps were easier for humans to classify than control masks in a speeded recognition experiment. Similarly, we find that recognition performance in the same ANN models was likewise influenced by masking input images using human visual selectivity maps. This work contributes a new approach to evaluating the biological and psychological validity of leading ANNs as models of human vision: by examining their similarities and differences in terms of their visual selectivity to the information contained in images.
****

GRIN: Generative Relation and Intention Network for Multi-agent Trajectory Prediction

Longyuan Li, Jian Yao, Li Wenliang, Tong He, Tianjun Xiao, Junchi Yan, David Wipf, Zheng Zhang

Learning the distribution of future trajectories conditioned on the past is a crucial problem for understanding multi-agent systems. This is challenging because humans make decisions based on complex social relations and personal intents, resulting in highly complex uncertainties over trajectories. To address this problem, we propose a conditional deep generative model that combines advances in graph neural networks. The prior and recognition model encodes two types of latent codes for each agent: an inter-agent latent code to represent social relations and an intra-agent latent code to represent agent intentions. The decoder is carefully devised to leverage the codes in a disentangled way to predict multi-modal future trajectory distribution. Specifically, a graph attention network built upon inter-agent latent code is used to learn continuous pair-wise relations, and an agent's motion is controlled by its latent intents and its observations of all other agents. Through experiments on both synthetic and real-world datasets, we show that our model outperforms previous work in multiple performance metrics. We also show that our model generates realistic multi-modal trajectories.
****

Instance-Dependent Partial Label Learning

Ning Xu, Congyu Qiao, Xin Geng, Min-Ling Zhang

Partial label learning (PLL) is a typical weakly supervised learning problem, where each training example is associated with a set of candidate labels among which only one is true. Most existing PLL approaches assume that the incorrect labe

ls in each training example are randomly picked as the candidate labels. However, this assumption is not realistic since the candidate labels are always instance-dependent. In this paper, we consider instance-dependent PLL and assume that each example is associated with a latent label distribution constituted by the real number of each label, representing the degree to each label describing the feature. The incorrect label with a high degree is more likely to be annotated as the candidate label. Therefore, the latent label distribution is the essential labeling information in partially labeled examples and worth being leveraged for predictive model training. Motivated by this consideration, we propose a novel PLL method that recovers the label distribution as a label enhancement (LE) process and trains the predictive model iteratively in every epoch. Specifically, we assume the true posterior density of the latent label distribution takes on the variational approximate Dirichlet density parameterized by an inference model. Then the evidence lower bound is deduced for optimizing the inference model and the label distributions generated from the variational posterior are utilized for training the predictive model. Experiments on benchmark and real-world datasets validate the effectiveness of the proposed method. Source code is available at https://github.com/palm-ml/valen.
****

Deep Learning with Label Differential Privacy

Badih Ghazi, Noah Golowich, Ravi Kumar, Pasin Manurangsi, Chiyuan Zhang

The Randomized Response (RR) algorithm is a classical technique to improve robustness in survey aggregation, and has been widely adopted in applications with differential privacy guarantees. We propose a novel algorithm, Randomized Response with Prior (RRWithPrior), which can provide more accurate results while maintaining the same level of privacy guaranteed by RR. We then apply RRWithPrior to learn neural networks with label differential privacy (LabelDP), and show that when only the label needs to be protected, the model performance can be significantly improved over the previous state-of-the-art private baselines. Moreover, we study different ways to obtain priors, which when used with RRWithPrior can additionally improve the model performance, further reducing the accuracy gap between private and non-private models. We complement the empirical results with theoretical analysis showing that LabelDP is provably easier than protecting both the inputs and labels.
****

Semialgebraic Representation of Monotone Deep Equilibrium Models and Applications to Certification

Tong Chen, Jean B. Lasserre, Victor Magron, Edouard Pauwels

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 Role of Global Labels in Few-Shot Classification and How to Infer Them

Ruohan Wang, Massimiliano Pontil, Carlo Ciliberto

Few-shot learning is a central problem in meta-learning, where learners must quickly adapt to new tasks given limited training data. Recently, feature pre-training has become a ubiquitous component in state-of-the-art meta-learning methods and is shown to provide significant performance improvement. However, there is limited theoretical understanding of the connection between pre-training and meta-learning. Further, pre-training requires global labels shared across tasks, which may be unavailable in practice. In this paper, we show why exploiting pre-training is theoretically advantageous for meta-learning, and in particular the critical role of global labels. This motivates us to propose Meta Label Learning (MeLa), a novel meta-learning framework that automatically infers global labels to obtains robust few-shot models. Empirically, we demonstrate that MeLa is competitive with existing methods and provide extensive ablation experiments to highlight its key properties.
****

NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Recon

struction
Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, Wenping Wang
We present a novel neural surface reconstruction method, called NeuS, for reconstructing objects and scenes with high fidelity from 2D image inputs. Existing neural surface reconstruction approaches, such as DVR [Niemeyer et al., 2020] and IDR [Yariv et al., 2020], require foreground mask as supervision, easily get trapped in local minima, and therefore struggle with the reconstruction of objects with severe self-occlusion or thin structures. Meanwhile, recent neural methods for novel view synthesis, such as NeRF [Mildenhall et al., 2020] and its variants, use volume rendering to produce a neural scene representation with robustness of optimization, even for highly complex objects. However, extracting high-quality surfaces from this learned implicit representation is difficult because there are not sufficient surface constraints in the representation. In NeuS, we propose to represent a surface as the zero-level set of a signed distance function (SDF) and develop a new volume rendering method to train a neural SDF representation. We observe that the conventional volume rendering method causes inherent geometric errors (i.e. bias) for surface reconstruction, and therefore propose a new formulation that is free of bias in the first order of approximation, thus leading to more accurate surface reconstruction even without the mask supervision. Experiments on the DTU dataset and the BlendedMVS dataset show that NeuS outperforms the state-of-the-arts in high-quality surface reconstruction, especially for objects and scenes with complex structures and self-occlusion.
****
Improved Guarantees for Offline Stochastic Matching via new Ordered Contention Resolution Schemes
Brian Brubach, Nathaniel Grammel, Will Ma, Aravind Srinivasan
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
UFC-BERT: Unifying Multi-Modal Controls for Conditional Image Synthesis
Zhu Zhang, Jianxin Ma, Chang Zhou, Rui Men, Zhikang Li, Ming Ding, Jie Tang, Jingren Zhou, Hongxia Yang
Conditional image synthesis aims to create an image according to some multi-modal guidance in the forms of textual descriptions, reference images, and image blocks to preserve, as well as their combinations. In this paper, instead of investigating these control signals separately, we propose a new two-stage architecture, UFC-BERT, to unify any number of multi-modal controls. In UFC-BERT, both the diverse control signals and the synthesized image are uniformly represented as a sequence of discrete tokens to be processed by Transformer. Different from existing two-stage autoregressive approaches such as DALL-E and VQGAN, UFC-BERT adopts non-autoregressive generation (NAR) at the second stage to enhance the holistic consistency of the synthesized image, to support preserving specified image blocks, and to improve the synthesis speed. Further, we design a progressive algorithm that iteratively improves the non-autoregressively generated image, with the help of two estimators developed for evaluating the compliance with the controls and evaluating the fidelity of the synthesized image, respectively. Extensive experiments on a newly collected large-scale clothing dataset M2C-Fashion and a facial dataset Multi-Modal CelebA-HQ verify that UFC-BERT can synthesize high-fidelity images that comply with flexible multi-modal controls.
****
Is Bang-Bang Control All You Need? Solving Continuous Control with Bernoulli Policies
Tim Seyde, Igor Gilitschenski, Wilko Schwarting, Bartolomeo Stellato, Martin Riedmiller, Markus Wulfmeier, Daniela Rus
Reinforcement learning (RL) for continuous control typically employs distributions whose support covers the entire action space. In this work, we investigate the colloquially known phenomenon that trained agents often prefer actions at the boundaries of that space. We draw theoretical connections to the emergence of ba

ng-bang behavior in optimal control, and provide extensive empirical evaluation across a variety of recent RL algorithms. We replace the normal Gaussian by a Bernoulli distribution that solely considers the extremes along each action dimension - a bang-bang controller. Surprisingly, this achieves state-of-the-art performance on several continuous control benchmarks - in contrast to robotic hardware, where energy and maintenance cost affect controller choices. Since exploration, learning, and the final solution are entangled in RL, we provide additional imitation learning experiments to reduce the impact of exploration on our analysis. Finally, we show that our observations generalize to environments that aim to model real-world challenges and evaluate factors to mitigate the emergence of bang-bang solutions. Our findings emphasise challenges for benchmarking continuous control algorithms, particularly in light of potential real-world applications.
****
Improving Generalization in Meta-RL with Imaginary Tasks from Latent Dynamics Mixture
Suyoung Lee, Sae-Young Chung
The generalization ability of most meta-reinforcement learning (meta-RL) methods is largely limited to test tasks that are sampled from the same distribution used to sample training tasks. To overcome the limitation, we propose Latent Dynamics Mixture (LDM) that trains a reinforcement learning agent with imaginary tasks generated from mixtures of learned latent dynamics. By training a policy on mixture tasks along with original training tasks, LDM allows the agent to prepare for unseen test tasks during training and prevents the agent from overfitting the training tasks. LDM significantly outperforms standard meta-RL methods in test returns on the gridworld navigation and MuJoCo tasks where we strictly separate the training task distribution and the test task distribution.
****
Localization with Sampling-Argmax
Jiefeng Li, Tong Chen, Ruiqi Shi, Yujing Lou, Yong-Lu Li, Cewu Lu
Soft-argmax operation is commonly adopted in detection-based methods to localize the target position in a differentiable manner. However, training the neural network with soft-argmax makes the shape of the probability map unconstrained. Consequently, the model lacks pixel-wise supervision through the map during training, leading to performance degradation. In this work, we propose sampling-argmax, a differentiable training method that imposes implicit constraints to the shape of the probability map by minimizing the expectation of the localization error. To approximate the expectation, we introduce a continuous formulation of the output distribution and develop a differentiable sampling process. The expectation can be approximated by calculating the average error of all samples drawn from the output distribution. We show that sampling-argmax can seamlessly replace the conventional soft-argmax operation on various localization tasks. Comprehensive experiments demonstrate the effectiveness and flexibility of the proposed method. Code is available at https://github.com/Jeff-sjtu/sampling-argmax
****
Improved Regularization and Robustness for Fine-tuning in Neural Networks
Dongyue Li, Hongyang Zhang
A widely used algorithm for transfer learning is fine-tuning, where a pre-trained model is fine-tuned on a target task with a small amount of labeled data. When the capacity of the pre-trained model is much larger than the size of the target data set, fine-tuning is prone to overfitting and "memorizing" the training labels. Hence, an important question is to regularize fine-tuning and ensure its robustness to noise. To address this question, we begin by analyzing the generalization properties of fine-tuning. We present a PAC-Bayes generalization bound that depends on the distance traveled in each layer during fine-tuning and the noise stability of the fine-tuned model. We empirically measure these quantities. Based on the analysis, we propose regularized self-labeling---the interpolation between regularization and self-labeling methods, including (i) layer-wise regularization to constrain the distance traveled in each layer; (ii) self label-correction and label-reweighting to correct mislabeled data points (that the model is

confident) and reweight less confident data points. We validate our approach on an extensive collection of image and text data sets using multiple pre-trained model architectures. Our approach improves baseline methods by 1.76% (on average) for seven image classification tasks and 0.75% for a few-shot classification task. When the target data set includes noisy labels, our approach outperforms baseline methods by 3.56% on average in two noisy settings.
****
BARTScore: Evaluating Generated Text as Text Generation
Weizhe Yuan, Graham Neubig, Pengfei Liu
A wide variety of NLP applications, such as machine translation, summarization, and dialog, involve text generation. One major challenge for these applications is how to evaluate whether such generated texts are actually fluent, accurate, or effective. In this work, we conceptualize the evaluation of generated text as a text generation problem, modeled using pre-trained sequence-to-sequence models. The general idea is that models trained to convert the generated text to/from a reference output or the source text will achieve higher scores when the generated text is better. We operationalize this idea using BART, an encoder-decoder based pre-trained model, and propose a metric BARTScore with a number of variants that can be flexibly applied in an unsupervised fashion to evaluation of text from different perspectives (e.g. informativeness, fluency, or factuality). BARTScore is conceptually simple and empirically effective. It can outperform existing top-scoring metrics in 16 of 22 test settings, covering evaluation of 16 datasets (e.g., machine translation, text summarization) and 7 different perspectives (e.g., informativeness, factuality). Code to calculate BARTScore is available at https://github.com/neulab/BARTScore, and we have released an interactive leaderboard for meta-evaluation at http://explainaboard.nlpedia.ai/leaderboard/task-meval/ on the ExplainaBoard platform, which allows us to interactively understand the strengths, weaknesses, and complementarity of each metric.
****
An analysis of Ermakov-Zolotukhin quadrature using kernels
Ayoub Belhadji
We study a quadrature, proposed by Ermakov and Zolotukhin in the sixties, through the lens of kernel methods. The nodes of this quadrature rule follow the distribution of a determinantal point process, while the weights are defined through a linear system, similarly to the optimal kernel quadrature. In this work, we show how these two classes of quadrature are related, and we prove a tractable formula of the expected value of the squared worst-case integration error on the unit ball of an RKHS of the former quadrature. In particular, this formula involves the eigenvalues of the corresponding kernel and leads to improving on the existing theoretical guarantees of the optimal kernel quadrature with determinantal point processes.
****
Towards Understanding Why Lookahead Generalizes Better Than SGD and Beyond
Pan Zhou, Hanshu Yan, Xiaotong Yuan, Jiashi Feng, Shuicheng Yan
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 Market Equilibrium with Application to Fair Division
Yuan Gao, Alex Peysakhovich, Christian Kroer
Computing market equilibria is a problem of both theoretical and applied interest. Much research to date focuses on the case of static Fisher markets with full information on buyers' utility functions and item supplies. Motivated by real-world markets, we consider an online setting: individuals have linear, additive utility functions; items arrive sequentially and must be allocated and priced irrevocably. We define the notion of an online market equilibrium in such a market as time-indexed allocations and prices which guarantee buyer optimality and market clearance in hindsight. We propose a simple, scalable and interpretable allocation and pricing dynamics termed as PACE. When items are drawn i.i.d. from an un

known distribution (with a possibly continuous support), we show that PACE leads to an online market equilibrium asymptotically. In particular, PACE ensures that buyers' time-averaged utilities converge to the equilibrium utilities w.r.t. a static market with item supplies being the unknown distribution and that buyers' time-averaged expenditures converge to their per-period budget. Hence, many desirable properties of market equilibrium-based fair division such as envy-freeness, Pareto optimality, and the proportional-share guarantee are also attained asymptotically in the online setting. Next, we extend the dynamics to handle quasilinear buyer utilities, which gives the first online algorithm for computing first-price pacing equilibria. Finally, numerical experiments on real and synthetic datasets show that the dynamics converges quickly under various metrics.
****
Dynamic Resolution Network
Mingjian Zhu, Kai Han, Enhua Wu, Qiulin Zhang, Ying Nie, Zhenzhong Lan, Yunhe Wang
Deep convolutional neural networks (CNNs) are often of sophisticated design with numerous learnable parameters for the accuracy reason. To alleviate the expensive costs of deploying them on mobile devices, recent works have made huge efforts for excavating redundancy in pre-defined architectures. Nevertheless, the redundancy on the input resolution of modern CNNs has not been fully investigated, i.e., the resolution of input image is fixed. In this paper, we observe that the smallest resolution for accurately predicting the given image is different using the same neural network. To this end, we propose a novel dynamic-resolution network (DRNet) in which the input resolution is determined dynamically based on each input sample. Wherein, a resolution predictor with negligible computational costs is explored and optimized jointly with the desired network. Specifically, the predictor learns the smallest resolution that can retain and even exceed the original recognition accuracy for each image. During the inference, each input image will be resized to its predicted resolution for minimizing the overall computation burden. We then conduct extensive experiments on several benchmark networks and datasets. The results show that our DRNet can be embedded in any off-the-shelf network architecture to obtain a considerable reduction in computational complexity. For instance, DR-ResNet-50 achieves similar performance with an about 34% computation reduction, while gaining 1.4% accuracy increase with 10% computation reduction compared to the original ResNet-50 on ImageNet. Code will be available at https://gitee.com/mindspore/models/tree/master/research/cv/DRNet.
****
Gauge Equivariant Transformer
Lingshen He, Yiming Dong, Yisen Wang, Dacheng Tao, Zhouchen Lin
Attention mechanism has shown great performance and efficiency in a lot of deep learning models, in which relative position encoding plays a crucial role. However, when introducing attention to manifolds, there is no canonical local coordinate system to parameterize neighborhoods. To address this issue, we propose an equivariant transformer to make our model agnostic to the orientation of local coordinate systems (\textit{i.e.}, gauge equivariant), which employs multi-head self-attention to jointly incorporate both position-based and content-based information. To enhance expressive ability, we adopt regular field of cyclic groups as feature fields in intermediate layers, and propose a novel method to parallel transport the feature vectors in these fields. In addition, we project the position vector of each point onto its local coordinate system to disentangle the orientation of the coordinate system in ambient space (\textit{i.e.}, global coordinate system), achieving rotation invariance. To the best of our knowledge, we are the first to introduce gauge equivariance to self-attention, thus name our model Gauge Equivariant Transformer (GET), which can be efficiently implemented on triangle meshes. Extensive experiments show that GET achieves state-of-the-art performance on two common recognition tasks.
****
Unsupervised Object-Based Transition Models For 3D Partially Observable Environments
Antonia Creswell, Rishabh Kabra, Chris Burgess, Murray Shanahan

We present a slot-wise, object-based transition model that decomposes a scene in
to objects, aligns them (with respect to a slot-wise object memory) to maintain
a consistent order across time, and predicts how those objects evolve over succe
ssive frames. The model is trained end-to-end without supervision using transiti
on losses at the level of the object-structured representation rather than pixel
s. Thanks to the introduction of our novel alignment module, the model deals pro
perly with two issues that are not handled satisfactorily by other transition mo
dels, namely object persistence and object identity. We show that the combinatio
n of an object-level loss and correct object alignment over time enables the mod
el to outperform a state-of-the-art baseline, and allows it to deal well with ob
ject occlusion and re-appearance in partially observable environments.
****
Robust Contrastive Learning Using Negative Samples with Diminished Semantics
Songwei Ge, Shlok Mishra, Chun-Liang Li, Haohan Wang, David Jacobs
Unsupervised learning has recently made exceptional progress because of the deve
lopment of more effective contrastive learning methods. However, CNNs are prone
to depend on low-level features that humans deem non-semantic. This dependency h
as been conjectured to induce a lack of robustness to image perturbations or dom
ain shift. In this paper, we show that by generating carefully designed negative
 samples, contrastive learning can learn more robust representations with less d
ependence on such features. Contrastive learning utilizes positive pairs which p
reserve semantic information while perturbing superficial features in the traini
ng images. Similarly, we propose to generate negative samples in a reversed way,
 where only the superfluous instead of the semantic features are preserved. We d
evelop two methods, texture-based and patch-based augmentations, to generate neg
ative samples. These samples achieve better generalization, especially under ou
t-of-domain settings. We also analyze our method and the generated texture-based
 samples, showing that texture features are indispensable in classifying particu
lar ImageNet classes and especially finer classes. We also show that the model b
ias between texture and shape features favors them differently under different t
est settings.
****
General Low-rank Matrix Optimization: Geometric Analysis and Sharper Bounds
Haixiang Zhang, Yingjie Bi, Javad Lavaei
Requests for name changes in the electronic proceedings will be accepted with no
 questions asked. However name changes may cause bibliographic tracking issues.
  Authors are asked to consider this carefully and discuss it with their co-auth
ors prior to requesting a name change in the electronic proceedings.
****
Flow Network based Generative Models for Non-Iterative Diverse Candidate Generat
ion
Emmanuel Bengio, Moksh Jain, Maksym Korablyov, Doina Precup, Yoshua Bengio
This paper is about the problem of learning a stochastic policy for generating a
n object (like a molecular graph) from a sequence of actions, such that the prob
ability of generating an object is proportional to a given positive reward for t
hat object. Whereas standard return maximization tends to converge to a single r
eturn-maximizing sequence, there are cases where we would like to sample a diver
se set of high-return solutions. These arise, for example, in black-box function
 optimization when few rounds are possible, each with large batches of queries,
where the batches should be diverse, e.g., in the design of new molecules. One c
an also see this as a problem of approximately converting an energy function to
a generative distribution. While MCMC methods can achieve that, they are expensi
ve and generally only perform local exploration. Instead, training a generative
policy amortizes the cost of search during training and yields to fast generatio
n. Using insights from Temporal Difference learning, we propose GFlowNet, based
 on a view of the generative process as a flow network, making it possible to ha
ndle the tricky case where different trajectories can yield the same final state
, e.g., there are many ways to sequentially add atoms to generate some molecular
 graph. We cast the set of trajectories as a flow and convert the flow consisten
cy equations into a learning objective, akin to the casting of the Bellman equat

ions into Temporal Difference methods. We prove that any global minimum of the proposed objectives yields a policy which samples from the desired distribution, and demonstrate the improved performance and diversity of GFlowNet on a simple domain where there are many modes to the reward function, and on a molecule synthesis task.
****
Policy Finetuning: Bridging Sample-Efficient Offline and Online Reinforcement Learning
Tengyang Xie, Nan Jiang, Huan Wang, Caiming Xiong, Yu Bai
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Reducing Information Bottleneck for Weakly Supervised Semantic Segmentation
Jungbeom Lee, Jooyoung Choi, Jisoo Mok, Sungroh Yoon
Weakly supervised semantic segmentation produces pixel-level localization from class labels; however, a classifier trained on such labels is likely to focus on a small discriminative region of the target object. We interpret this phenomenon using the information bottleneck principle: the final layer of a deep neural network, activated by the sigmoid or softmax activation functions, causes an information bottleneck, and as a result, only a subset of the task-relevant information is passed on to the output. We first support this argument through a simulated toy experiment and then propose a method to reduce the information bottleneck by removing the last activation function. In addition, we introduce a new pooling method that further encourages the transmission of information from non-discriminative regions to the classification. Our experimental evaluations demonstrate that this simple modification significantly improves the quality of localization maps on both the PASCAL VOC 2012 and MS COCO 2014 datasets, exhibiting a new state-of-the-art performance for weakly supervised semantic segmentation.
****
SGD: The Role of Implicit Regularization, Batch-size and Multiple-epochs
Ayush Sekhari, Karthik Sridharan, Satyen Kale
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
AC-GC: Lossy Activation Compression with Guaranteed Convergence
R David Evans, Tor Aamodt
Parallel hardware devices (e.g., graphics processor units) have limited high-bandwidth memory capacity.This negatively impacts the training of deep neural networks (DNNs) by increasing runtime and/or decreasing accuracy when reducing model and/or batch size to fit this capacity. Lossy compression is a promising approach to tackling memory capacity constraints, but prior approaches rely on hyperparameter search to achieve a suitable trade-off between convergence and compression, negating runtime benefits. In this paper we build upon recent developments on Stochastic Gradient Descent convergence to prove an upper bound on the expected loss increase when training with compressed activation storage. We then express activation compression error in terms of this bound, allowing the compression rate to adapt to training conditions automatically. The advantage of our approach, called AC-GC, over existing lossy compression frameworks is that, given a preset allowable increase in loss, significant compression without significant increase in error can be achieved with a single training run. When combined with error-bounded methods, AC-GC achieves 15.1x compression with an average accuracy change of 0.1% on text and image datasets. AC-GC functions on any model composed of the layers analyzed and, by avoiding compression rate search, reduces overall training time by 4.6x over SuccessiveHalving.
****
Label Noise SGD Provably Prefers Flat Global Minimizers

Alex Damian, Tengyu Ma, Jason D. Lee
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Can we have it all? On the Trade-off between Spatial and Adversarial Robustness of Neural Networks
Sandesh Kamath, Amit Deshpande, Subrahmanyam Kambhampati Venkata, Vineeth N Balasubramanian
(Non-)robustness of neural networks to small, adversarial pixel-wise perturbations, and as more recently shown, to even random spatial transformations (e.g., translations, rotations) entreats both theoretical and empirical understanding. Spatial robustness to random translations and rotations is commonly attained via equivariant models (e.g., StdCNNs, GCNNs) and training augmentation, whereas adversarial robustness is typically achieved by adversarial training. In this paper, we prove a quantitative trade-off between spatial and adversarial robustness in a simple statistical setting. We complement this empirically by showing that: (a) as the spatial robustness of equivariant models improves by training augmentation with progressively larger transformations, their adversarial robustness worsens progressively, and (b) as the state-of-the-art robust models are adversarially trained with progressively larger pixel-wise perturbations, their spatial robustness drops progressively. Towards achieving Pareto-optimality in this trade-off, we propose a method based on curriculum learning that trains gradually on more difficult perturbations (both spatial and adversarial) to improve spatial and adversarial robustness simultaneously.
****
Universal Off-Policy Evaluation
Yash Chandak, Scott Niekum, Bruno da Silva, Erik Learned-Miller, Emma Brunskill, Philip S. Thomas
When faced with sequential decision-making problems, it is often useful to be able to predict what would happen if decisions were made using a new policy. Those predictions must often be based on data collected under some previously used decision-making rule. Many previous methods enable such off-policy (or counterfactual) estimation of the expected value of a performance measure called the return. In this paper, we take the first steps towards a 'universal off-policy estimator' (UnO)---one that provides off-policy estimates and high-confidence bounds for any parameter of the return distribution. We use UnO for estimating and simultaneously bounding the mean, variance, quantiles/median, inter-quantile range, CVaR, and the entire cumulative distribution of returns. Finally, we also discuss UnO's applicability in various settings, including fully observable, partially observable (i.e., with unobserved confounders), Markovian, non-Markovian, stationary, smoothly non-stationary, and discrete distribution shifts.
****
A Non-commutative Extension of Lee-Seung's Algorithm for Positive Semidefinite Factorizations
Yong Sheng Soh, Antonios Varvitsiotis
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Efficiently Identifying Task Groupings for Multi-Task Learning
Chris Fifty, Ehsan Amid, Zhe Zhao, Tianhe Yu, Rohan Anil, Chelsea Finn
Multi-task learning can leverage information learned by one task to benefit the training of other tasks. Despite this capacity, naively training all tasks together in one model often degrades performance, and exhaustively searching through combinations of task groupings can be prohibitively expensive. As a result, efficiently identifying the tasks that would benefit from training together remains a challenging design question without a clear solution. In this paper, we sugges

t an approach to select which tasks should train together in multi-task learning models. Our method determines task groupings in a single run by training all tasks together and quantifying the effect to which one task's gradient would affect another task's loss. On the large-scale Taskonomy computer vision dataset, we find this method can decrease test loss by 10.0% compared to simply training all tasks together while operating 11.6 times faster than a state-of-the-art task grouping method.
****
Instance-Conditioned GAN
Arantxa Casanova, Marlene Careil, Jakob Verbeek, Michal Drozdzal, Adriana Romero Soriano
Generative Adversarial Networks (GANs) can generate near photo realistic images in narrow domains such as human faces. Yet, modeling complex distributions of datasets such as ImageNet and COCO-Stuff remains challenging in unconditional settings. In this paper, we take inspiration from kernel density estimation techniques and introduce a non-parametric approach to modeling distributions of complex datasets. We partition the data manifold into a mixture of overlapping neighborhoods described by a datapoint and its nearest neighbors, and introduce a model, called instance-conditioned GAN (IC-GAN), which learns the distribution around each datapoint. Experimental results on ImageNet and COCO-Stuff show that IC-GAN significantly improves over unconditional models and unsupervised data partitioning baselines. Moreover, we show that IC-GAN can effortlessly transfer to datasets not seen during training by simply changing the conditioning instances, and still generate realistic images. Finally, we extend IC-GAN to the class-conditional case and show semantically controllable generation and competitive quantitative results on ImageNet; while improving over BigGAN on ImageNet-LT. Code and trained models to reproduce the reported results are available at https://github.com/facebookresearch/ic_gan.
****
DeepSITH: Efficient Learning via Decomposition of What and When Across Time Scales
Brandon Jacques, Zoran Tiganj, Marc Howard, Per B Sederberg
Extracting temporal relationships over a range of scales is a hallmark ofhuman perception and cognition---and thus it is a critical feature of machinelearning applied to real-world problems. Neural networks are either plaguedby the exploding/vanishing gradient problem in recurrent neural networks(RNNs) or must adjust their parameters to learn the relevant time scales(e.g., in LSTMs). This paper introduces DeepSITH, a deep network comprisingbiologically-inspired Scale-Invariant Temporal History (SITH) modules inseries with dense connections between layers. Each SITH module is simply aset of time cells coding what happened when with a geometrically-spaced set oftime lags. The dense connections between layers change the definition of whatfrom one layer to the next. The geometric series of time lags implies thatthe network codes time on a logarithmic scale, enabling DeepSITH network tolearn problems requiring memory over a wide range of time scales. We compareDeepSITH to LSTMs and other recent RNNs on several time series prediction anddecoding tasks. DeepSITH achieves results comparable to state-of-the-artperformance on these problems and continues to perform well even as the delays are increased.
****
A Gaussian Process-Bayesian Bernoulli Mixture Model for Multi-Label Active Learning
Weishi Shi, Dayou Yu, Qi 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.
****
Differentially Private Empirical Risk Minimization under the Fairness Lens
Cuong Tran, My Dinh, Ferdinando Fioretto
Differential Privacy (DP) is an important privacy-enhancing technology for priva

te machine learning systems. It allows to measure and bound the risk associated with an individual participation in a computation. However, it was recently observed that DP learning systems may exacerbate bias and unfairness for different groups of individuals. This paper builds on these important observations and sheds light on the causes of the disparate impacts arising in the problem of differentially private empirical risk minimization. It focuses on the accuracy disparity arising among groups of individuals in two well-studied DP learning methods: output perturbation and differentially private stochastic gradient descent. The paper analyzes which data and model properties are responsible for the disproportionate impacts, why these aspects are affecting different groups disproportionately, and proposes guidelines to mitigate these effects. The proposed approach is evaluated on several datasets and settings.
****
A Unified View of cGANs with and without Classifiers
Si-An Chen, Chun-Liang Li, Hsuan-Tien Lin
Conditional Generative Adversarial Networks (cGANs) are implicit generative models which allow to sample from class-conditional distributions. Existing cGANs are based on a wide range of different discriminator designs and training objectives. One popular design in earlier works is to include a classifier during training with the assumption that good classifiers can help eliminate samples generated with wrong classes. Nevertheless, including classifiers in cGANs often comes with a side effect of only generating easy-to-classify samples. Recently, some representative cGANs avoid the shortcoming and reach state-of-the-art performance without having classifiers. Somehow it remains unanswered whether the classifiers can be resurrected to design better cGANs. In this work, we demonstrate that classifiers can be properly leveraged to improve cGANs. We start by using the decomposition of the joint probability distribution to connect the goals of cGANs and classification as a unified framework. The framework, along with a classic energy model to parameterize distributions, justifies the use of classifiers for cGANs in a principled manner. It explains several popular cGAN variants, such as ACGAN, ProjGAN, and ContraGAN, as special cases with different levels of approximations, which provides a unified view and brings new insights to understanding cGANs. Experimental results demonstrate that the design inspired by the proposed framework outperforms state-of-the-art cGANs on multiple benchmark datasets, especially on the most challenging ImageNet. The code is available at https://github.com/sian-chen/PyTorch-ECGAN.
****
Online and Offline Reinforcement Learning by Planning with a Learned Model
Julian Schrittwieser, Thomas Hubert, Amol Mandhane, Mohammadamin Barekatain, Ioannis Antonoglou, David Silver
Learning efficiently from small amounts of data has long been the focus of model-based reinforcement learning, both for the online case when interacting with the environment, and the offline case when learning from a fixed dataset. However, to date no single unified algorithm could demonstrate state-of-the-art results for both settings.In this work, we describe the Reanalyse algorithm, which uses model-based policy and value improvement operators to compute improved training targets for existing data points, allowing for efficient learning at data budgets varying by several orders of magnitude. We further show that Reanalyse can also be used to learn completely without environment interactions, as in the case of Offline Reinforcement Learning (Offline RL). Combining Reanalyse with the MuZero algorithm, we introduce MuZero Unplugged, a single unified algorithm for any data budget, including Offline RL. In contrast to previous work, our algorithm requires no special adaptations for the off-policy or Offline RL settings. MuZero Unplugged sets new state-of-the-art results for Atari in the standard 200 million frame online setting as well as in the RL Unplugged Offline RL benchmark.
****
Stochastic Multi-Armed Bandits with Control Variates
Arun Verma, Manjesh Kumar Hanawal
This paper studies a new variant of the stochastic multi-armed bandits problem where auxiliary information about the arm rewards is available in the form of con

trol variates. In many applications like queuing and wireless networks, the arm rewards are functions of some exogenous variables. The mean values of these vari ables are known a priori from historical data and can be used as control variate s. Leveraging the theory of control variates, we obtain mean estimates with sma ller variance and tighter confidence bounds. We develop an upper confidence boun d based algorithm named UCB-CV and characterize the regret bounds in terms of th e correlation between rewards and control variates when they follow a multivaria te normal distribution. We also extend UCB-CV to other distributions using resam pling methods like Jackknifing and Splitting. Experiments on synthetic problem i nstances validate performance guarantees of the proposed algorithms.
****
Near-Optimal No-Regret Learning in General Games
Constantinos Daskalakis, Maxwell Fishelson, Noah Golowich
Requests for name changes in the electronic proceedings will be accepted with 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.
****
Improving Self-supervised Learning with Automated Unsupervised Outlier Arbitrati on
Yu Wang, Jingyang Lin, Jingjing Zou, Yingwei Pan, Ting Yao, Tao Mei
Our work reveals a structured shortcoming of the existing mainstream self-superv ised learning methods. Whereas self-supervised learning frameworks usually take the prevailing perfect instance level invariance hypothesis for granted, we care fully investigate the pitfalls behind. Particularly, we argue that the existing augmentation pipeline for generating multiple positive views naturally introduce s out-of-distribution (OOD) samples that undermine the learning of the downstrea m tasks. Generating diverse positive augmentations on the input does not always pay off in benefiting downstream tasks. To overcome this inherent deficiency, we introduce a lightweight latent variable model UOTA, targeting the view sampling issue for self-supervised learning. UOTA adaptively searches for the most impor tant sampling region to produce views, and provides viable choice for outlier-ro bust self-supervised learning approaches. Our method directly generalizes to man y mainstream self-supervised learning approaches, regardless of the loss's natur e contrastive or not. We empirically show UOTA's advantage over the state-of-the -art self-supervised paradigms with evident margin, which well justifies the exi stence of the OOD sample issue embedded in the existing approaches. Especially, we theoretically prove that the merits of the proposal boil down to guaranteed e stimator variance and bias reduction. Code is available: https://github.com/ssl-codelab/uota.
****
Improving Anytime Prediction with Parallel Cascaded Networks and a Temporal-Diff erence Loss
Michael Iuzzolino, Michael C. Mozer, Samy Bengio
Although deep feedforward neural networks share some characteristics with the pr imate visual system, a key distinction is their dynamics. Deep nets typically o perate in serial stages wherein each layer completes its computation before proc essing begins in subsequent layers.   In contrast, biological systems have casca ded dynamics: information propagates from neurons at all layers in parallel but transmission occurs gradually over time, leading to speed-accuracy trade offs ev en in feedforward architectures. We explore the consequences of biologically ins pired parallel hardware by constructing cascaded ResNets in which each residual block has propagation delays but all blocks update in parallel in a stateful man ner. Because information transmitted through skip connections avoids delays, the functional depth of the architecture increases over time, yielding anytime pre dictions that improve with internal-processing time. We introduce a temporal-dif ference training loss that achieves a strictly superior speed-accuracy profile o ver standard losses and enables the cascaded architecture to outperform state-of -the-art anytime-prediction methods. The cascaded architecture has intriguing pr operties, including: it classifies typical instances more rapidly than atypical

instances; it is more robust to both persistent and transient noise than is a conventional ResNet; and its time-varying output trace provides a signal that can be exploited to improve information processing and inference.
****

Identifiable Generative models for Missing Not at Random Data Imputation
Chao Ma, Cheng Zhang
Real-world datasets often have missing values associated with complex generative processes, where the cause of the missingness may not be fully observed. This is known as missing not at random (MNAR) data. However, many imputation methods do not take into account the missingness mechanism, resulting in biased imputation values when MNAR data is present. Although there are a few methods that have considered the MNAR scenario, their model's identifiability under MNAR is generally not guaranteed. That is, model parameters can not be uniquely determined even with infinite data samples, hence the imputation results given by such models can still be biased. This issue is especially overlooked by many modern deep generative models. In this work, we fill in this gap by systematically analyzing the identifiability of generative models under MNAR. Furthermore, we propose a practical deep generative model which can provide identifiability guarantees under mild assumptions, for a wide range of MNAR mechanisms. Our method demonstrates a clear advantage for tasks on both synthetic data and multiple real-world scenarios with MNAR data.
****

DNN-based Topology Optimisation: Spatial Invariance and Neural Tangent Kernel
Benjamin Dupuis, Arthur Jacot
We study the Solid Isotropic Material Penalization (SIMP) method with a density field generated by a fully-connected neural network, taking the coordinates as inputs. In the large width limit, we show that the use of DNNs leads to a filtering effect similar to traditional filtering techniques for SIMP, with a filter described by the Neural Tangent Kernel (NTK). This filter is however not invariant under translation, leading to visual artifacts and non-optimal shapes. We propose two embeddings of the input coordinates, which lead to (approximate) spatial invariance of the NTK and of the filter. We empirically confirm our theoretical observations and study how the filter size is affected by the architecture of the network. Our solution can easily be applied to any other coordinates-based generation method.
****

Baleen: Robust Multi-Hop Reasoning at Scale via Condensed Retrieval
Omar Khattab, Christopher Potts, Matei Zaharia
Multi-hop reasoning (i.e., reasoning across two or more documents) is a key ingredient for NLP models that leverage large corpora to exhibit broad knowledge. To retrieve evidence passages, multi-hop models must contend with a fast-growing search space across the hops, represent complex queries that combine multiple information needs, and resolve ambiguity about the best order in which to hop between training passages. We tackle these problems via Baleen, a system that improves the accuracy of multi-hop retrieval while learning robustly from weak training signals in the many-hop setting. To tame the search space, we propose condensed retrieval, a pipeline that summarizes the retrieved passages after each hop into a single compact context. To model complex queries, we introduce a focused late interaction retriever that allows different parts of the same query representation to match disparate relevant passages. Lastly, to infer the hopping dependencies among unordered training passages, we devise latent hop ordering, a weak-supervision strategy in which the trained retriever itself selects the sequence of hops. We evaluate Baleen on retrieval for two-hop question answering and many-hop claim verification, establishing state-of-the-art performance.
****

Local Hyper-Flow Diffusion
Kimon Fountoulakis, Pan Li, Shenghao 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.
****
Permuton-induced Chinese Restaurant Process
Masahiro Nakano, Yasuhiro Fujiwara, Akisato Kimura, Takeshi Yamada, naonori ueda
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Faster Algorithms and Constant Lower Bounds for the Worst-Case Expected Error
Jonah Brown-Cohen
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 Learning Domain-Invariant Representations for Transfer Learning with Multiple Sources
Trung Phung, Trung Le, Tung-Long Vuong, Toan Tran, Anh Tran, Hung Bui, Dinh Phung
Domain adaptation (DA) benefits from the rigorous theoretical works that study its insightful characteristics and various aspects, e.g., learning domain-invariant representations and its trade-off. However, it seems not the case for the multiple source DA and domain generalization (DG) settings which are remarkably more complicated and sophisticated due to the involvement of multiple source domains and potential unavailability of target domain during training. In this paper, we develop novel upper-bounds for the target general loss which appeal us to define two kinds of domain-invariant representations. We further study the pros and cons as well as the trade-offs of enforcing learning each domain-invariant representation. Finally, we conduct experiments to inspect the trade-off of these representations for offering practical hints regarding how to use them in practice and explore other interesting properties of our developed theory.
****
You Never Cluster Alone
Yuming Shen, Ziyi Shen, Menghan Wang, Jie Qin, Philip Torr, Ling Shao
Recent advances in self-supervised learning with instance-level contrastive objectives facilitate unsupervised clustering. However, a standalone datum is not perceiving the context of the holistic cluster, and may undergo sub-optimal assignment. In this paper, we extend the mainstream contrastive learning paradigm to a cluster-level scheme, where all the data subjected to the same cluster contribute to a unified representation that encodes the context of each data group. Contrastive learning with this representation then rewards the assignment of each datum. To implement this vision, we propose twin-contrast clustering (TCC). We define a set of categorical variables as clustering assignment confidence, which links the instance-level learning track with the cluster-level one. On one hand, with the corresponding assignment variables being the weight, a weighted aggregation along the data points implements the set representation of a cluster. We further propose heuristic cluster augmentation equivalents to enable cluster-level contrastive learning. On the other hand, we derive the evidence lower-bound of the instance-level contrastive objective with the assignments. By reparametrizing the assignment variables, TCC is trained end-to-end, requiring no alternating steps. Extensive experiments show that TCC outperforms the state-of-the-art on benchmarked datasets.
****
Dynamic COVID risk assessment accounting for community virus exposure from a spatial-temporal transmission model
Yuan Chen, Wenbo Fei, Qinxia Wang, Donglin Zeng, Yuanjia Wang
COVID-19 pandemic has caused unprecedented negative impacts on our society, including further exposing inequity and disparity in public health. To study the impact of socioeconomic factors on COVID transmission, we first propose a spatial-t

emporal model to examine the socioeconomic heterogeneity and spatial correlation of COVID-19 transmission at the community level. Second, to assess the individual risk of severe COVID-19 outcomes after a positive diagnosis, we propose a dynamic, varying-coefficient model that integrates individual-level risk factors from electronic health records (EHRs) with community-level risk factors. The underlying neighborhood prevalence of infections (both symptomatic and pre-symptomatic) predicted from the previous spatial-temporal model is included in the individual risk assessment so as to better capture the background risk of virus exposure for each individual. We design a weighting scheme to mitigate multiple selection biases inherited in EHRs of COVID patients. We analyze COVID transmission data in New York City (NYC, the epicenter of the first surge in the United States) and EHRs from NYC hospitals, where time-varying effects of community risk factors and significant interactions between individual- and community-level risk factors are detected. By examining the socioeconomic disparity of infection risks and interaction among the risk factors, our methods can assist public health decision-making and facilitate better clinical management of COVID patients.
****
Dueling Bandits with Adversarial Sleeping
Aadirupa Saha, 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.
****
Beware of the Simulated DAG! Causal Discovery Benchmarks May Be Easy to Game
Alexander Reisach, Christof Seiler, Sebastian Weichwald
Simulated DAG models may exhibit properties that, perhaps inadvertently, render their structure identifiable and unexpectedly affect structure learning algorithms. Here, we show that marginal variance tends to increase along the causal order for generically sampled additive noise models. We introduce varsortability as a measure of the agreement between the order of increasing marginal variance and the causal order. For commonly sampled graphs and model parameters, we show that the remarkable performance of some continuous structure learning algorithms can be explained by high varsortability and matched by a simple baseline method. Yet, this performance may not transfer to real-world data where varsortability may be moderate or dependent on the choice of measurement scales. On standardized data, the same algorithms fail to identify the ground-truth DAG or its Markov equivalence class. While standardization removes the pattern in marginal variance, we show that data generating processes that incur high varsortability also leave a distinct covariance pattern that may be exploited even after standardization. Our findings challenge the significance of generic benchmarks with independently drawn parameters. The code is available at https://github.com/Scriddie/Varsortability.
****
Automated Dynamic Mechanism Design
Hanrui Zhang, Vincent Conitzer
We study Bayesian automated mechanism design in unstructured dynamic environments, where a principal repeatedly interacts with an agent, and takes actions based on the strategic agent's report of the current state of the world. Both the principal and the agent can have arbitrary and potentially different valuations for the actions taken, possibly also depending on the actual state of the world. Moreover, at any time, the state of the world may evolve arbitrarily depending on the action taken by the principal. The goal is to compute an optimal mechanism which maximizes the principal's utility in the face of the self-interested strategic agent.We give an efficient algorithm for computing optimal mechanisms, with or without payments, under different individual-rationality constraints, when the time horizon is constant. Our algorithm is based on a sophisticated linear program formulation, which can be customized in various ways to accommodate richer constraints. For environments with large time horizons, we show that the principal's optimal utility is hard to approximate within a certain constant facto

r, complementing our algorithmic result. These results paint a relatively compl
ete picture for automated dynamic mechanism design in unstructured environments.
  We further consider a special case of the problem where the agent is myopic, a
nd give a refined efficient algorithm whose time complexity scales linearly in t
he time horizon.   In the full version of the paper, we show that memoryless me
chanisms, which are without loss of generality optimal in Markov decision proces
ses without strategic behavior, do not provide a good solution for our problem,
in terms of both optimality and computational tractability. Moreover, we presen
t experimental results where our algorithms are applied to synthetic dynamic env
ironments with different characteristics, which not only serve as a proof of con
cept for our algorithms, but also exhibit intriguing phenomena in dynamic mechan
ism design.
****
A generative nonparametric Bayesian model for whole genomes
Alan Amin, Eli N Weinstein, Debora Marks
Generative probabilistic modeling of biological sequences has widespread existin
g and potential use across biology and biomedicine, particularly given advances
in high-throughput sequencing, synthesis and editing. However, we still lack met
hods with nucleotide resolution that are tractable at the scale of whole genomes
 and that can achieve high predictive accuracy in theory and practice. In this a
rticle we propose a new generative sequence model, the Bayesian embedded autoreg
ressive (BEAR) model, which uses a parametric autoregressive model to specify a
conjugate prior over a nonparametric Bayesian Markov model. We explore, theoreti
cally and empirically, applications of BEAR models to a variety of statistical p
roblems including density estimation, robust parameter estimation, goodness-of-f
it tests, and two-sample tests. We prove rigorous asymptotic consistency results
 including nonparametric posterior concentration rates. We scale inference in BE
AR models to datasets containing tens of billions of nucleotides. On genomic, tr
anscriptomic, and metagenomic sequence data we show that BEAR models provide lar
ge increases in predictive performance as compared to parametric autoregressive
models, among other results. BEAR models offer a flexible and scalable framework
, with theoretical guarantees, for building and critiquing generative models at
the whole genome scale.
****
Robust Predictable Control
Ben Eysenbach, Russ R. Salakhutdinov, Sergey Levine
Requests for name changes in the electronic proceedings will be accepted with 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.
****
Unsupervised Speech Recognition
Alexei Baevski, Wei-Ning Hsu, Alexis CONNEAU, Michael Auli
Despite rapid progress in the recent past, current speech recognition systems st
ill require labeled training data which limits this technology to a small fracti
on of the languages spoken around the globe. This paper describes wav2vec-U, sho
rt for wav2vec Unsupervised, a method to train speech recognition models without
 any labeled data. We leverage self-supervised speech representations to segment
 unlabeled audio and learn a mapping from these representations to phonemes via
adversarial training. The right representations are key to the success of our me
thod. Compared to the best previous unsupervised work, wav2vec-U reduces the pho
ne error rate on the TIMIT benchmark from 26.1 to 11.3. On the larger English Li
brispeech benchmark, wav2vec-U achieves a word error rate of 5.9 on test-other,
rivaling some of the best published systems trained on 960 hours of labeled data
 from only two years ago. We also experiment on nine other languages, including
low-resource languages such as Kyrgyz, Swahili and Tatar.
****
Robustness between the worst and average case
Leslie Rice, Anna Bair, Huan Zhang, J. Zico Kolter
Several recent works in machine learning have focused on evaluating the test-tim

e robustness of a classifier: how well the classifier performs not just on the t
arget domain it was trained upon, but upon perturbed examples. In these setting
s, the focus has largely been on two extremes of robustness: the robustness to p
erturbations drawn at random from within some distribution (i.e., robustness to
random perturbations), and the robustness to the worst case perturbation in some
 set (i.e., adversarial robustness). In this paper, we argue that a sliding sca
le between these two extremes provides a valuable additional metric by which to
gauge robustness. Specifically, we illustrate that each of these two extremes is
 naturally characterized by a (functional) q-norm over perturbation space, with
q=1 corresponding to robustness to random perturbations and q=\infty correspondi
ng to adversarial perturbations. We then present the main technical contributio
n of our paper: a method for efficiently estimating the value of these norms by
interpreting them as the partition function of a particular distribution, then u
sing path sampling with MCMC methods to estimate this partition function (either
 traditional Metropolis-Hastings for non-differentiable perturbations, or Hamilt
onian Monte Carlo for differentiable perturbations). We show that our approach
provides substantially better estimates than simple random sampling of the actua
l “intermediate-q” robustness of both standard, data-augmented, and adversariall
y-trained classifiers, illustrating a clear tradeoff between classifiers that op
timize different metrics. Code for reproducing experiments can be found at https
://github.com/locuslab/intermediate_robustness.
****
Online Learning and Control of Complex Dynamical Systems from Sensory Input
Oumayma Bounou, Jean Ponce, Justin Carpentier
Identifying an effective model of a dynamical system from sensory data and using
 it for future state prediction and control is challenging. Recent data-driven a
lgorithms based on Koopman theory are a promising approach to this problem, but
they typically never update the model once it has been identified from a relativ
ely small set of observation, thus making long-term prediction and control diffi
cult for realistic systems, in robotics or fluid mechanics for example. This pap
er introduces a novel method for learning an embedding of the state space with l
inear dynamics from sensory data. Unlike previous approaches, the dynamics model
 can be updated online and thus easily applied to systems with non-linear dynami
cs in the original configuration space. The proposed approach is evaluated empi
rically on several classical dynamical systems and sensory modalities, with good
 performance on long-term prediction and control.
****
Self-Supervised Bug Detection and Repair
Miltiadis Allamanis, Henry Jackson-Flux, Marc Brockschmidt
Machine learning-based program analyses have recently shown the promise of integ
rating formal and probabilistic reasoning towards aiding software development. H
owever, in the absence of large annotated corpora, training these analyses is ch
allenging. Towards addressing this, we present BugLab, an approach for self-supe
rvised learning of bug detection and repair. BugLab co-trains two models: (1) a
 detector model that learns to detect and repair bugs in code, (2) a selector mo
del that learns to create buggy code for the detector to use as training data. A
 Python implementation of BugLab improves by 30% upon baseline methods on a test
 dataset of 2374 real-life bugs and finds 19 previously unknown bugs in open-sou
rce software.
****
Faster Neural Network Training with Approximate Tensor Operations
Menachem Adelman, Kfir Levy, Ido Hakimi, Mark Silberstein
We propose a novel technique for faster deep neural network training which syste
matically applies sample-based approximation to the constituent tensor operation
s, i.e., matrix multiplications and convolutions. We introduce new sampling tech
niques, study their theoretical properties, and prove that they provide the same
 convergence guarantees when applied to SGD training. We apply approximate tenso
r operations to single and multi-node training of MLP and CNN networks on MNIST,
 CIFAR-10 and ImageNet datasets. We demonstrate up to 66% reduction in the amoun
t of computations and communication, and up to 1.37x faster training time while

maintaining negligible or no impact on the final test accuracy.
****

Learning Interpretable Decision Rule Sets: A Submodular Optimization Approach

Fan Yang, Kai He, Linxiao Yang, Hongxia Du, Jingbang Yang, Bo Yang, Liang Sun

Rule sets are highly interpretable logical models in which the predicates for decision are expressed in disjunctive normal form (DNF, OR-of-ANDs), or, equivalently, the overall model comprises an unordered collection of if-then decision rules. In this paper, we consider a submodular optimization based approach for learning rule sets. The learning problem is framed as a subset selection task in which a subset of all possible rules needs to be selected to form an accurate and interpretable rule set. We employ an objective function that exhibits submodularity and thus is amenable to submodular optimization techniques. To overcome the difficulty arose from dealing with the exponential-sized ground set of rules, the subproblem of searching a rule is casted as another subset selection task that asks for a subset of features. We show it is possible to write the induced objective function for the subproblem as a difference of two submodular (DS) functions to make it approximately solvable by DS optimization algorithms. Overall, the proposed approach is simple, scalable, and likely to be benefited from further research on submodular optimization. Experiments on real datasets demonstrate the effectiveness of our method.
****

Spatial-Temporal Super-Resolution of Satellite Imagery via Conditional Pixel Synthesis

Yutong He, Dingjie Wang, Nicholas Lai, William Zhang, Chenlin Meng, Marshall Burke, David Lobell, Stefano Ermon

High-resolution satellite imagery has proven useful for a broad range of tasks, including measurement of global human population, local economic livelihoods, and biodiversity, among many others. Unfortunately, high-resolution imagery is both infrequently collected and expensive to purchase, making it hard to efficiently and effectively scale these downstream tasks over both time and space. We propose a new conditional pixel synthesis model that uses abundant, low-cost, low-resolution imagery to generate accurate high-resolution imagery at locations and times in which it is unavailable. We show that our model attains photo-realistic sample quality and outperforms competing baselines on a key downstream task – object counting – particularly in geographic locations where conditions on the ground are changing rapidly.
****

On Memorization in Probabilistic Deep Generative Models

Gerrit van den Burg, Chris Williams

Recent advances in deep generative models have led to impressive results in a variety of application domains. Motivated by the possibility that deep learning models might memorize part of the input data, there have been increased efforts to understand how memorization arises. In this work, we extend a recently proposed measure of memorization for supervised learning (Feldman, 2019) to the unsupervised density estimation problem and adapt it to be more computationally efficient. Next, we present a study that demonstrates how memorization can occur in probabilistic deep generative models such as variational autoencoders. This reveals that the form of memorization to which these models are susceptible differs fundamentally from mode collapse and overfitting. Furthermore, we show that the proposed memorization score measures a phenomenon that is not captured by commonly-used nearest neighbor tests. Finally, we discuss several strategies that can be used to limit memorization in practice. Our work thus provides a framework for understanding problematic memorization in probabilistic generative models.
****

You Are the Best Reviewer of Your Own Papers: An Owner-Assisted Scoring Mechanism

Weijie Su

I consider the setting where reviewers offer very noisy scores for a number of items for the selection of high-quality ones (e.g., peer review of large conference proceedings) whereas the owner of these items knows the true underlying score

s but prefers not to provide this information. To address this withholding of in formation, in this paper, I introduce the Isotonic Mechanism, a simple and effic ient approach to improving on the imprecise raw scores by leveraging certain inf ormation that the owner is incentivized to provide. This mechanism takes as inpu t the ranking of the items from best to worst provided by the owner, in addition to the raw scores provided by the reviewers. It reports adjusted scores for the items by solving a convex optimization problem. Under certain conditions, I sho w that the owner's optimal strategy is to honestly report the true ranking of th e items to her best knowledge in order to maximize the expected utility. Moreove r, I prove that the adjusted scores provided by this owner-assisted mechanism ar e indeed significantly moreaccurate than the raw scores provided by the reviewer s. This paper concludes with several extensions of the Isotonic Mechanism and so me refinements of the mechanism for practical considerations.
****

Garment4D: Garment Reconstruction from Point Cloud Sequences

Fangzhou Hong, Liang Pan, Zhongang Cai, Ziwei Liu

Learning to reconstruct 3D garments is important for dressing 3D human bodies of different shapes in different poses. Previous works typically rely on 2D images as input, which however suffer from the scale and pose ambiguities. To circumve nt the problems caused by 2D images, we propose a principled framework, Garment4 D, that uses 3D point cloud sequences of dressed humans for garment reconstructi on. Garment4D has three dedicated steps: sequential garments registration, canon ical garment estimation, and posed garment reconstruction. The main challenges a re two-fold: 1) effective 3D feature learning for fine details, and 2) capture o f garment dynamics caused by the interaction between garments and the human body , especially for loose garments like skirts. To unravel these problems, we intro duce a novel Proposal-Guided Hierarchical Feature Network and Iterative Graph Co nvolution Network, which integrate both high-level semantic features and low-lev el geometric features for fine details reconstruction. Furthermore, we propose a Temporal Transformer for smooth garment motions capture. Unlike non-parametric methods, the reconstructed garment meshes by our method are separable from the h uman body and have strong interpretability, which is desirable for downstream ta sks. As the first attempt at this task, high-quality reconstruction results are qualitatively and quantitatively illustrated through extensive experiments. Code s are available at https://github.com/hongfz16/Garment4D.
****

Fast Policy Extragradient Methods for Competitive Games with Entropy Regularizat ion

Shicong Cen, Yuting Wei, Yuejie Chi

This paper investigates the problem of computing the equilibrium of competitive games, which is often modeled as a constrained saddle-point optimization problem with probability simplex constraints. Despite recent efforts in understanding t he last-iterate convergence of extragradient methods in the unconstrained settin g, the theoretical underpinnings of these methods in the constrained settings, e specially those using multiplicative updates, remain highly inadequate, even whe n the objective function is bilinear. Motivated by the algorithmic role of entro py regularization in single-agent reinforcement learning and game theory, we dev elop provably efficient extragradient methods to find the quantal response equil ibrium (QRE)---which are solutions to zero-sum two-player matrix games with entr opy regularization---at a linear rate. The proposed algorithms can be implemente d in a decentralized manner, where each player executes symmetric and multiplica tive updates iteratively using its own payoff without observing the opponent's a ctions directly. In addition, by controlling the knob of entropy regularization , the proposed algorithms can locate an approximate Nash equilibrium of the unre gularized matrix game at a sublinear rate without assuming the Nash equilibrium to be unique. Our methods also lead to efficient policy extragradient algorithms for solving entropy-regularized zero-sum Markov games at a linear rate. All of our convergence rates are nearly dimension-free, which are independent of the si ze of the state and action spaces up to logarithm factors, highlighting the posi tive role of entropy regularization for accelerating convergence.

****
Shift-Robust GNNs: Overcoming the Limitations of Localized Graph Training data
Qi Zhu, Natalia Ponomareva, Jiawei Han, Bryan Perozzi
There has been a recent surge of interest in designing Graph Neural Networks (GNNs) for semi-supervised learning tasks. Unfortunately this work has assumed that the nodes labeled for use in training were selected uniformly at random (i.e. are an IID sample). However in many real world scenarios gathering labels for graph nodes is both expensive and inherently biased -- so this assumption can not be met. GNNs can suffer poor generalization when this occurs, by overfitting to superfluous regularities present in the training data. In this work we present a method, Shift-Robust GNN (SR-GNN), designed to account for distributional differences between biased training data and the graph's true inference distribution. SR-GNN adapts GNN models for the presence of distributional shifts between the nodes which have had labels provided for training and the rest of the dataset. We illustrate the effectiveness of SR-GNN in a variety of experiments with biased training datasets on common GNN benchmark datasets for semi-supervised learning, where we see that SR-GNN outperforms other GNN baselines by accuracy, eliminating at least (~40%) of the negative effects introduced by biased training data. On the largest dataset we consider, ogb-arxiv, we observe an 2% absolute improvement over the baseline and reduce 30% of the negative effects.
****
RIM: Reliable Influence-based Active Learning on Graphs
Wentao Zhang, Yexin Wang, Zhenbang You, Meng Cao, Ping Huang, Jiulong Shan, Zhi Yang, Bin CUI
Message passing is the core of most graph models such as Graph Convolutional Network (GCN) and Label Propagation (LP), which usually require a large number of clean labeled data to smooth out the neighborhood over the graph. However, the labeling process can be tedious, costly, and error-prone in practice. In this paper, we propose to unify active learning (AL) and message passing towards minimizing labeling costs, e.g., making use of few and unreliable labels that can be obtained cheaply. We make two contributions towards that end. First, we open up a perspective by drawing a connection between AL enforcing message passing and social influence maximization, ensuring that the selected samples effectively improve the model performance. Second, we propose an extension to the influence model that incorporates an explicit quality factor to model label noise. In this way, we derive a fundamentally new AL selection criterion for GCN and LP--reliable influence maximization (RIM)--by considering quantity and quality of influence simultaneously. Empirical studies on public datasets show that RIM significantly outperforms current AL methods in terms of accuracy and efficiency.
****
Dynamical Wasserstein Barycenters for Time-series Modeling
Kevin Cheng, Shuchin Aeron, Michael C. Hughes, Eric L Miller
Many time series can be modeled as a sequence of segments representing high-level discrete states, such as running and walking in a human activity application. Flexible models should describe the system state and observations in stationary ``pure-state'' periods as well as transition periods between adjacent segments, such as a gradual slowdown between running and walking. However, most prior work assumes instantaneous transitions between pure discrete states. We propose a dynamical Wasserstein barycentric (DWB) model that estimates the system state over time as well as the data-generating distributions of pure states in an unsupervised manner. Our model assumes each pure state generates data from a multivariate normal distribution, and characterizes transitions between states via displacement-interpolation specified by the Wasserstein barycenter. The system state is represented by a barycentric weight vector which evolves over time via a random walk on the simplex. Parameter learning leverages the natural Riemannian geometry of Gaussian distributions under the Wasserstein distance, which leads to improved convergence speeds. Experiments on several human activity datasets show that our proposed DWB model accurately learns the generating distribution of pure states while improving state estimation for transition periods compared to the commonly used linear interpolation mixture models.

****
RelaySum for Decentralized Deep Learning on Heterogeneous Data
Thijs Vogels, Lie He, Anastasiia Koloskova, Sai Praneeth Karimireddy, Tao Lin, Sebastian U. Stich, Martin Jaggi
In decentralized machine learning, workers compute model updates on their local data.Because the workers only communicate with few neighbors without central coordination, these updates propagate progressively over the network.This paradigm enables distributed training on networks without all-to-all connectivity, helping to protect data privacy as well as to reduce the communication cost of distributed training in data centers.A key challenge, primarily in decentralized deep learning, remains the handling of differences between the workers' local data distributions.To tackle this challenge, we introduce the RelaySum mechanism for information propagation in decentralized learning.RelaySum uses spanning trees to distribute information exactly uniformly across all workers with finite delays depending on the distance between nodes.In contrast, the typical gossip averaging mechanism only distributes data uniformly asymptotically while using the same communication volume per step as RelaySum.We prove that RelaySGD, based on this mechanism, is independent of data heterogeneity and scales to many workers, enabling highly accurate decentralized deep learning on heterogeneous data.
****
Transformers Generalize DeepSets and Can be Extended to Graphs & Hypergraphs
Jinwoo Kim, Saeyoon Oh, Seunghoon 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.
****
No Regrets for Learning the Prior in Bandits
Soumya Basu, Branislav Kveton, Manzil Zaheer, Csaba Szepesvari
We propose AdaTS, a Thompson sampling algorithm that adapts sequentially to bandit tasks that it interacts with. The key idea in AdaTS is to adapt to an unknown task prior distribution by maintaining a distribution over its parameters. When solving a bandit task, that uncertainty is marginalized out and properly accounted for. AdaTS is a fully-Bayesian algorithm that can be implemented efficiently in several classes of bandit problems. We derive upper bounds on its Bayes regret that quantify the loss due to not knowing the task prior, and show that it is small. Our theory is supported by experiments, where AdaTS outperforms prior algorithms and works well even in challenging real-world problems.
****
Encoding Robustness to Image Style via Adversarial Feature Perturbations
Manli Shu, Zuxuan Wu, Micah Goldblum, Tom Goldstein
Adversarial training is the industry standard for producing models that are robust to small adversarial perturbations. However, machine learning practitioners need models that are robust to other kinds of changes that occur naturally, such as changes in the style or illumination of input images. Such changes in input distribution have been effectively modeled as shifts in the mean and variance of deep image features. We adapt adversarial training by directly perturbing feature statistics, rather than image pixels, to produce models that are robust to various unseen distributional shifts. We explore the relationship between these perturbations and distributional shifts by visualizing adversarial features. Our proposed method, Adversarial Batch Normalization (AdvBN), is a single network layer that generates worst-case feature perturbations during training. By fine-tuning neural networks on adversarial feature distributions, we observe improved robustness of networks to various unseen distributional shifts, including style variations and image corruptions. In addition, we show that our proposed adversarial feature perturbation can be complementary to existing image space data augmentation methods, leading to improved performance. The source code and pre-trained models are released at \url{https://github.com/azshue/AdvBN}.
****
Continuized Accelerations of Deterministic and Stochastic Gradient Descents, and

of Gossip Algorithms
Mathieu Even, Raphaël Berthier, Francis Bach, Nicolas Flammarion, Hadrien Hendrikx, Pierre Gaillard, Laurent Massoulié, Adrien Taylor
We introduce the ``continuized'' Nesterov acceleration, a close variant of Nesterov acceleration whose variables are indexed by a continuous time parameter. The two variables continuously mix following a linear ordinary differential equation and take gradient steps at random times. This continuized variant benefits from the best of the continuous and the discrete frameworks: as a continuous process, one can use differential calculus to analyze convergence and obtain analytical expressions for the parameters; but a discretization of the continuized process can be computed exactly with convergence rates similar to those of Nesterov original acceleration. We show that the discretization has the same structure as Nesterov acceleration, but with random parameters. We provide continuized Nesterov acceleration under deterministic as well as stochastic gradients, with either additive or multiplicative noise. Finally, using our continuized framework and expressing the gossip averaging problem as the stochastic minimization of a certain energy function, we provide the first rigorous acceleration of asynchronous gossip algorithms.
****
Natural continual learning: success is a journey, not (just) a destination
Ta-Chu Kao, Kristopher Jensen, Gido van de Ven, Alberto Bernacchia, Guillaume Hennequin
Biological agents are known to learn many different tasks over the course of their lives, and to be able to revisit previous tasks and behaviors with little to no loss in performance. In contrast, artificial agents are prone to ‘catastrophic forgetting’ whereby performance on previous tasks deteriorates rapidly as new ones are acquired. This shortcoming has recently been addressed using methods that encourage parameters to stay close to those used for previous tasks. This can be done by (i) using specific parameter regularizers that map out suitable destinations in parameter space, or (ii) guiding the optimization journey by projecting gradients into subspaces that do not interfere with previous tasks. However, these methods often exhibit subpar performance in both feedforward and recurrent neural networks, with recurrent networks being of interest to the study of neural dynamics supporting biological continual learning. In this work, we propose Natural Continual Learning (NCL), a new method that unifies weight regularization and projected gradient descent. NCL uses Bayesian weight regularization to encourage good performance on all tasks at convergence and combines this with gradient projection using the prior precision, which prevents catastrophic forgetting during optimization. Our method outperforms both standard weight regularization techniques and projection based approaches when applied to continual learning problems in feedforward and recurrent networks. Finally, the trained networks evolve task-specific dynamics that are strongly preserved as new tasks are learned, similar to experimental findings in biological circuits.
****
Individual Privacy Accounting via a Rényi Filter
Vitaly Feldman, Tijana Zrnic
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 Quantization for Vision Transformer
Zhenhua Liu, Yunhe Wang, Kai Han, Wei Zhang, Siwei Ma, Wen Gao
Recently, transformer has achieved remarkable performance on a variety of computer vision applications. Compared with mainstream convolutional neural networks, vision transformers are often of sophisticated architectures for extracting powerful feature representations, which are more difficult to be developed on mobile devices. In this paper, we present an effective post-training quantization algorithm for reducing the memory storage and computational costs of vision transformers. Basically, the quantization task can be regarded as finding the optimal lo

w-bit quantization intervals for weights and inputs, respectively. To preserve t
he functionality of the attention mechanism, we introduce a ranking loss into th
e conventional quantization objective that aims to keep the relative order of th
e self-attention results after quantization. Moreover, we thoroughly analyze the
 relationship between quantization loss of different layers and the feature dive
rsity, and explore a mixed-precision quantization scheme by exploiting the nucle
ar norm of each attention map and output feature. The effectiveness of the propo
sed method is verified on several benchmark models and datasets, which outperfor
ms the state-of-the-art post-training quantization algorithms. For instance, we
can obtain an 81.29% top-1 accuracy using DeiT-B model on ImageNet dataset with
about 8-bit quantization. Code will be available at https://gitee.com/mindspore/
models/tree/master/research/cv/VT-PTQ.
****
Unsupervised Part Discovery from Contrastive Reconstruction
Subhabrata Choudhury, Iro Laina, Christian Rupprecht, Andrea Vedaldi
The goal of self-supervised visual representation learning is to learn strong, t
ransferable image representations, with the majority of research focusing on obj
ect or scene level. On the other hand, representation learning at part level has
 received significantly less attention. In this paper, we propose an unsupervise
d approach to object part discovery and segmentation and make three contribution
s. First, we construct a proxy task through a set of objectives that encourages
the model to learn a meaningful decomposition of the image into its parts. Secon
dly, prior work argues for reconstructing or clustering pre-computed features as
 a proxy to parts; we show empirically that this alone is unlikely to find meani
ngful parts; mainly because of their low resolution and the tendency of classifi
cation networks to spatially smear out information. We suggest that image recons
truction at the level of pixels can alleviate this problem, acting as a compleme
ntary cue. Lastly, we show that the standard evaluation based on keypoint regres
sion does not correlate well with segmentation quality and thus introduce differ
ent metrics, NMI and ARI, that better characterize the decomposition of objects
into parts. Our method yields semantic parts which are consistent across fine-gr
ained but visually distinct categories, outperforming the state of the art on th
ree benchmark datasets. Code is available at the project page: https://www.robot
s.ox.ac.uk/~vgg/research/unsup-parts/.
****
ASSANet: An Anisotropic Separable Set Abstraction for Efficient Point Cloud Repr
esentation Learning
Guocheng Qian, Hasan Hammoud, Guohao Li, Ali Thabet, Bernard Ghanem
Requests for name changes in the electronic proceedings will be accepted with 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 Empirical Investigation of Domain Generalization with Empirical Risk Minimize
rs
Ramakrishna Vedantam, David Lopez-Paz, David J. Schwab
Recent work demonstrates that deep neural networks trained using Empirical Risk
Minimization (ERM) can generalize under distribution shift, outperforming specia
lized training algorithms for domain generalization. The goal of this paper is t
o further understand this phenomenon. In particular, we study the extent to whic
h the seminal domain adaptation theory of Ben-David et al. (2007) explains the p
erformance of ERMs. Perhaps surprisingly, we find that this theory does not prov
ide a tight explanation of the out-of-domain generalization observed across a la
rge number of ERM models trained on three popular domain generalization datasets
. This motivates us to investigate other possible measures—that, however, lack t
heory—which could explain generalization in this setting. Our investigation reve
als that measures relating to the Fisher information, predictive entropy, and ma
ximum mean discrepancy are good predictors of the out-of-distribution generaliza
tion of ERM models. We hope that our work helps galvanize the community towards
building a better understanding of when deep networks trained with ERM generaliz

e out-of-distribution.
****
Fair Sequential Selection Using Supervised Learning Models
Mohammad Mahdi Khalili, Xueru Zhang, Mahed Abroshan
We consider a selection problem where sequentially arrived applicants apply for a limited number of positions/jobs. At each time step, a decision maker accepts or rejects the given applicant using a pre-trained supervised learning model until all the vacant positions are filled. In this paper, we discuss whether the fairness notions (e.g., equal opportunity, statistical parity, etc.) that are commonly used in classification problems are suitable for the sequential selection problems. In particular, we show that even with a pre-trained model that satisfies the common fairness notions, the selection outcomes may still be biased against certain demographic groups. This observation implies that the fairness notions used in classification problems are not suitable for a selection problem where the applicants compete for a limited number of positions. We introduce a new fairness notion, ``Equal Selection (ES),'' suitable for sequential selection problems and propose a post-processing approach to satisfy the ES fairness notion. We also consider a setting where the applicants have privacy concerns, and the decision maker only has access to the noisy version of sensitive attributes. In this setting, we can show that the \textit{perfect} ES fairness can still be attained under certain conditions.
****
Towards Sample-efficient Overparameterized Meta-learning
Yue Sun, Adhyyan Narang, Ibrahim Gulluk, Samet Oymak, Maryam Fazel
An overarching goal in machine learning is to build a generalizable model with few samples. To this end, overparameterization has been the subject of immense interest to explain the generalization ability of deep nets even when the size of the dataset is smaller than that of the model. While the prior literature focuses on the classical supervised setting, this paper aims to demystify overparameterization for meta-learning. Here we have a sequence of linear-regression tasks and we ask: (1) Given earlier tasks, what is the optimal linear representation of features for a new downstream task? and (2) How many samples do we need to build this representation? This work shows that surprisingly, overparameterization arises as a natural answer to these fundamental meta-learning questions. Specifically, for (1), we first show that learning the optimal representation coincides with the problem of designing a task-aware regularization to promote inductive bias. We leverage this inductive bias to explain how the downstream task actually benefits from overparameterization, in contrast to prior works on few-shot learning. For (2), we develop a theory to explain how feature covariance can implicitly help reduce the sample complexity well below the degrees of freedom and lead to small estimation error. We then integrate these findings to obtain an overall performance guarantee for our meta-learning algorithm. Numerical experiments on real and synthetic data verify our insights on overparameterized meta-learning.
****
ScaleCert: Scalable Certified Defense against Adversarial Patches with Sparse Superficial Layers
Husheng Han, Kaidi Xu, Xing Hu, Xiaobing Chen, Ling LIANG, Zidong Du, Qi Guo, Yanzhi Wang, Yunji 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.
****
Towards mental time travel: a hierarchical memory for reinforcement learning agents
Andrew Lampinen, Stephanie Chan, Andrea Banino, Felix Hill
Reinforcement learning agents often forget details of the past, especially after delays or distractor tasks. Agents with common memory architectures struggle to recall and integrate across multiple timesteps of a past event, or even to reca

ll the details of a single timestep that is followed by distractor tasks. To add ress these limitations, we propose a Hierarchical Chunk Attention Memory (HCAM), that helps agents to remember the past in detail. HCAM stores memories by dividing the past into chunks, and recalls by first performing high-level attention over coarse summaries of the chunks, and then performing detailed attention within only the most relevant chunks. An agent with HCAM can therefore "mentally time-travel"--remember past events in detail without attending to all intervening events. We show that agents with HCAM substantially outperform agents with other memory architectures at tasks requiring long-term recall, retention, or reasoning over memory. These include recalling where an object is hidden in a 3D environment, rapidly learning to navigate efficiently in a new neighborhood, and rapidly learning and retaining new words. Agents with HCAM can extrapolate to task sequences much longer than they were trained on, and can even generalize zero-shot from a meta-learning setting to maintaining knowledge across episodes. HCAM improve agent sample efficiency, generalization, and generality (by solving tasks that previously required specialized architectures). Our work is a step towards agents that can learn, interact, and adapt in complex and temporally-extended environments.
****
Beyond Tikhonov: faster learning with self-concordant losses, via iterative regularization
Gaspard Beugnot, Julien Mairal, Alessandro Rudi
The theory of spectral filtering is a remarkable tool to understand the statistical properties of learning with kernels. For least squares, it allows to derive various regularization schemes that yield faster convergence rates of the excess risk than with Tikhonov regularization. This is typically achieved by leveraging classical assumptions called source and capacity conditions, which characterize the difficulty of the learning task. In order to understand estimators derived from other loss functions, Marteau-Ferey et al. have extended the theory of Tikhonov regularization to generalized self concordant loss functions (GSC), which contain, e.g., the logistic loss. In this paper, we go a step further and show that fast and optimal rates can be achieved for GSC by using the iterated Tikhonov regularization scheme, which is intrinsically related to the proximal point method in optimization, and overcomes the limitation of the classical Tikhonov regularization.
****
Variational Bayesian Reinforcement Learning with Regret Bounds
Brendan O'Donoghue
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Logarithmic Regret from Sublinear 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.
****
Independent mechanism analysis, a new concept?
Luigi Gresele, Julius von Kügelgen, Vincent Stimper, Bernhard Schölkopf, Michel Besserve
Independent component analysis provides a principled framework for unsupervised representation learning, with solid theory on the identifiability of the latent code that generated the data, given only observations of mixtures thereof. Unfortunately, when the mixing is nonlinear, the model is provably nonidentifiable, since statistical independence alone does not sufficiently constrain the problem. Identifiability can be recovered in settings where additional, typically observed variables are included in the generative process. We investigate an alternati

ve path and consider instead including assumptions reflecting the principle of i
ndependent causal mechanisms exploited in the field of causality. Specifically,
our approach is motivated by thinking of each source as independently influencin
g the mixing process. This gives rise to a framework which we term independent m
echanism analysis. We provide theoretical and empirical evidence that our approa
ch circumvents a number of nonidentifiability issues arising in nonlinear blind
source separation.
****
Momentum Centering and Asynchronous Update for Adaptive Gradient Methods
Juntang Zhuang, Yifan Ding, Tommy Tang, Nicha Dvornek, Sekhar C Tatikonda, James
 Duncan
Requests for name changes in the electronic proceedings will be accepted with 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.
****
Robustness via Uncertainty-aware Cycle Consistency
Uddeshya Upadhyay, Yanbei Chen, Zeynep Akata
Unpaired image-to-image translation refers to learning inter-image-domain mappin
g without corresponding image pairs. Existing methods learn deterministic mappin
gs without explicitly modelling the robustness to outliers or predictive uncerta
inty, leading to performance degradation when encountering unseen perturbations
at test time. To address this, we propose a novel probabilistic method based on
Uncertainty-aware Generalized Adaptive Cycle Consistency (UGAC), which models th
e per-pixel residual by generalized Gaussian distribution, capable of modelling
heavy-tailed distributions. We compare our model with a wide variety of state-of
-the-art methods on various challenging tasks including unpaired image translati
on of natural images spanning autonomous driving, maps, facades, and also in the
 medical imaging domain consisting of MRI. Experimental results demonstrate that
 our method exhibits stronger robustness towards unseen perturbations in test da
ta. Code is released here: https://github.com/ExplainableML/UncertaintyAwareCycl
eConsistency.
****
CBP: backpropagation with constraint on weight precision using a pseudo-Lagrange
 multiplier method
Guhyun Kim, Doo Seok Jeong
Backward propagation of errors (backpropagation) is a method to minimize objecti
ve functions (e.g., loss functions) of deep neural networks by identifying optim
al sets of weights and biases. Imposing constraints on weight precision is often
 required to alleviate prohibitive workloads on hardware. Despite the remarkable
 success of backpropagation, the algorithm itself is not capable of considering
such constraints unless additional algorithms are applied simultaneously. To add
ress this issue, we propose the constrained backpropagation (CBP) algorithm base
d on the pseudo-Lagrange multiplier method to obtain the optimal set of weights
that satisfy a given set of constraints. The defining characteristic of the prop
osed CBP algorithm is the utilization of a Lagrangian function (loss function pl
us constraint function) as its objective function. We considered various types o
f constraints — binary, ternary, one-bit shift, and two-bit shift weight constra
ints. As a post-training method, CBP applied to AlexNet, ResNet-18, ResNet-50, a
nd GoogLeNet on ImageNet, which were pre-trained using the conventional backprop
agation. For most cases, the proposed algorithm outperforms the state-of-the-art
 methods on ImageNet, e.g., 66.6\%, 74.4\%, and 64.0\% top-1 accuracy for ResNet
-18, ResNet-50, and GoogLeNet with binary weights, respectively. This highlights
 CBP as a learning algorithm to address diverse constraints with the minimal per
formance loss by employing appropriate constraint functions. The code for CBP is
 publicly available at \url{https://github.com/dooseokjeong/CBP}.
****
On the Sample Complexity of Privately Learning Axis-Aligned Rectangles
Menachem Sadigurschi, Uri Stemmer
Requests for name changes in the electronic proceedings will be accepted with no

questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Implicit Sparse Regularization: The Impact of Depth and Early Stopping
Jiangyuan Li, Thanh Nguyen, Chinmay Hegde, Ka Wai Wong
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 Generalization with Distributionally Robust Learning
Soumyadip Ghosh, Mark Squillante, Ebisa Wollega
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
No-regret Online Learning over Riemannian Manifolds
Xi Wang, Zhipeng Tu, Yiguang Hong, Yingyi Wu, Guodong Shi
We consider online optimization over Riemannian manifolds, where a learner attempts to minimize a sequence of time-varying loss functions defined on Riemannian manifolds. Though many Euclidean online convex optimization algorithms have been proven useful in a wide range of areas, less attention has been paid to their Riemannian counterparts. In this paper, we study Riemannian online gradient descent (R-OGD) on Hadamard manifolds for both geodesically convex and strongly geodesically convex loss functions, and Riemannian bandit algorithm (R-BAN) on Hadamard homogeneous manifolds for geodesically convex functions. We establish upper bounds on the regrets of the problem with respect to time horizon, manifold curvature, and manifold dimension. We also find a universal lower bound for the achievable regret by constructing an online convex optimization problem on Hadamard manifolds. All the obtained regret bounds match the corresponding results are provided in Euclidean spaces. Finally, some numerical experiments validate our theoretical results.
****
Landmark-Guided Subgoal Generation in Hierarchical Reinforcement Learning
Junsu Kim, Younggyo Seo, Jinwoo Shin
Goal-conditioned hierarchical reinforcement learning (HRL) has shown promising results for solving complex and long-horizon RL tasks. However, the action space of high-level policy in the goal-conditioned HRL is often large, so it results in poor exploration, leading to inefficiency in training. In this paper, we present HIerarchical reinforcement learning Guided by Landmarks (HIGL), a novel framework for training a high-level policy with a reduced action space guided by landmarks, i.e., promising states to explore. The key component of HIGL is twofold: (a) sampling landmarks that are informative for exploration and (b) encouraging the high level policy to generate a subgoal towards a selected landmark. For (a), we consider two criteria: coverage of the entire visited state space (i.e., dispersion of states) and novelty of states (i.e., prediction error of a state). For (b), we select a landmark as the very first landmark in the shortest path in a graph whose nodes are landmarks. Our experiments demonstrate that our framework outperforms prior-arts across a variety of control tasks, thanks to efficient exploration guided by landmarks.
****
Minimax Regret for Stochastic Shortest Path
Alon Cohen, Yonathan Efroni, Yishay Mansour, Aviv Rosenberg
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

Parametrized Quantum Policies for Reinforcement Learning
Sofiene Jerbi, Casper Gyurik, Simon Marshall, Hans Briegel, Vedran Dunjko
With the advent of real-world quantum computing, the idea that parametrized quantum computations can be used as hypothesis families in a quantum-classical machine learning system is gaining increasing traction. Such hybrid systems have already shown the potential to tackle real-world tasks in supervised and generative learning, and recent works have established their provable advantages in special artificial tasks. Yet, in the case of reinforcement learning, which is arguably most challenging and where learning boosts would be extremely valuable, no proposal has been successful in solving even standard benchmarking tasks, nor in showing a theoretical learning advantage over classical algorithms. In this work, we achieve both. We propose a hybrid quantum-classical reinforcement learning model using very few qubits, which we show can be effectively trained to solve several standard benchmarking environments. Moreover, we demonstrate, and formally prove, the ability of parametrized quantum circuits to solve certain learning tasks that are intractable to classical models, including current state-of-art deep neural networks, under the widely-believed classical hardness of the discrete logarithm problem.
****
On Pathologies in KL-Regularized Reinforcement Learning from Expert Demonstrations
Tim G. J. Rudner, Cong Lu, Michael A Osborne, Yarin Gal, Yee Teh
KL-regularized reinforcement learning from expert demonstrations has proved successful in improving the sample efficiency of deep reinforcement learning algorithms, allowing them to be applied to challenging physical real-world tasks. However, we show that KL-regularized reinforcement learning with behavioral reference policies derived from expert demonstrations can suffer from pathological training dynamics that can lead to slow, unstable, and suboptimal online learning. We show empirically that the pathology occurs for commonly chosen behavioral policy classes and demonstrate its impact on sample efficiency and online policy performance. Finally, we show that the pathology can be remedied by non-parametric behavioral reference policies and that this allows KL-regularized reinforcement learning to significantly outperform state-of-the-art approaches on a variety of challenging locomotion and dexterous hand manipulation tasks.
****
Conditional Generation Using Polynomial Expansions
Grigorios Chrysos, Markos Georgopoulos, Yannis Panagakis
Generative modeling has evolved to a notable field of machine learning. Deep polynomial neural networks (PNNs) have demonstrated impressive results in unsupervised image generation, where the task is to map an input vector (i.e., noise) to a synthesized image. However, the success of PNNs has not been replicated in conditional generation tasks, such as super-resolution. Existing PNNs focus on single-variable polynomial expansions which do not fare well to two-variable inputs, i.e., the noise variable and the conditional variable. In this work, we introduce a general framework, called CoPE, that enables a polynomial expansion of two input variables and captures their auto- and cross-correlations. We exhibit how CoPE can be trivially augmented to accept an arbitrary number of input variables. CoPE is evaluated in five tasks (class-conditional generation, inverse problems, edges-to-image translation, image-to-image translation, attribute-guided generation) involving eight datasets. The thorough evaluation suggests that CoPE can be useful for tackling diverse conditional generation tasks. The source code of CoPE is available at https://github.com/grigorisg9gr/polynomialnetsforconditionalgeneration.
****
Efficient constrained sampling via the mirror-Langevin algorithm
Kwangjun Ahn, Sinho Chewi
We propose a new discretization of the mirror-Langevin diffusion and give a crisp proof of its convergence. Our analysis uses relative convexity/smoothness and self-concordance, ideas which originated in convex optimization, together with a new result in optimal transport that generalizes the displacement convexity of

the entropy. Unlike prior works, our result both (1) requires much weaker assumptions on the mirror map and the target distribution, and (2) has vanishing bias as the step size tends to zero. In particular, for the task of sampling from a log-concave distribution supported on a compact set, our theoretical results are significantly better than the existing guarantees.
****

Adaptive Online Packing-guided Search for POMDPs
Chenyang Wu, Guoyu Yang, Zongzhang Zhang, Yang Yu, Dong Li, Wulong Liu, Jianye Hao
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

Turing Completeness of Bounded-Precision Recurrent Neural Networks
Stephen Chung, Hava Siegelmann
Previous works have proved that recurrent neural networks (RNNs) are Turing-complete. However, in the proofs, the RNNs allow for neurons with unbounded precision, which is neither practical in implementation nor biologically plausible. To remove this assumption, we propose a dynamically growing memory module made of neurons of fixed precision. The memory module dynamically recruits new neurons when more memories are needed, and releases them when memories become irrelevant. We prove that a 54-neuron bounded-precision RNN with growing memory modules can simulate a Universal Turing Machine, with time complexity linear in the simulated machine's time and independent of the memory size. The result is extendable to various other stack-augmented RNNs. Furthermore, we analyze the Turing completeness of both unbounded-precision and bounded-precision RNNs, revisiting and extending the theoretical foundations of RNNs.
****

End-to-end Multi-modal Video Temporal Grounding
Yi-Wen Chen, Yi-Hsuan Tsai, Ming-Hsuan Yang
We address the problem of text-guided video temporal grounding, which aims to identify the time interval of a certain event based on a natural language description. Different from most existing methods that only consider RGB images as visual features, we propose a multi-modal framework to extract complementary information from videos. Specifically, we adopt RGB images for appearance, optical flow for motion, and depth maps for image structure. While RGB images provide abundant visual cues of certain events, the performance may be affected by background clutters. Therefore, we use optical flow to focus on large motion and depth maps to infer the scene configuration when the action is related to objects recognizable with their shapes. To integrate the three modalities more effectively and enable inter-modal learning, we design a dynamic fusion scheme with transformers to model the interactions between modalities. Furthermore, we apply intra-modal self-supervised learning to enhance feature representations across videos for each modality, which also facilitates multi-modal learning. We conduct extensive experiments on the Charades-STA and ActivityNet Captions datasets, and show that the proposed method performs favorably against state-of-the-art approaches.
****

How Powerful are Performance Predictors in Neural Architecture Search?
Colin White, Arber Zela, Robin Ru, Yang Liu, Frank Hutter
Early methods in the rapidly developing field of neural architecture search (NAS) required fully training thousands of neural networks. To reduce this extreme computational cost, dozens of techniques have since been proposed to predict the final performance of neural architectures. Despite the success of such performance prediction methods, it is not well-understood how different families of techniques compare to one another, due to the lack of an agreed-upon evaluation metric and optimization for different constraints on the initialization time and query time. In this work, we give the first large-scale study of performance predictors by analyzing 31 techniques ranging from learning curve extrapolation, to weight-sharing, to supervised learning, to zero-cost proxies. We test a number of c

orrelation- and rank-based performance measures in a variety of settings, as wel
l as the ability of each technique to speed up predictor-based NAS frameworks. O
ur results act as recommendations for the best predictors to use in different se
ttings, and we show that certain families of predictors can be combined to achie
ve even better predictive power, opening up promising research directions. We re
lease our code, featuring a library of 31 performance predictors.
****
Stylized Dialogue Generation with Multi-Pass Dual Learning
Jinpeng Li, Yingce Xia, Rui Yan, Hongda Sun, Dongyan Zhao, Tie-Yan Liu
Stylized dialogue generation, which aims to generate a given-style response for
an input context, plays a vital role in intelligent dialogue systems. Considerin
g there is no parallel data between the contexts and the responses of target sty
le S1, existing works mainly use back translation to generate stylized synthetic
 data for training, where the data about context, target style S1 and an interme
diate style S0 is used. However, the interaction among these texts is not fully
exploited, and the pseudo contexts are not adequately modeled. To overcome the a
bove difficulties, we propose multi-pass dual learning (MPDL), which leverages t
he duality among the context, response of style S1 and response of style S_0. MP
DL builds mappings among the above three domains, where the context should be re
constructed by the MPDL framework, and the reconstruction error is used as the t
raining signal. To evaluate the quality of synthetic data, we also introduce dis
criminators that effectively measure how a pseudo sequence matches the specific
domain, and the evaluation result is used as the weight for that data. Evaluatio
n results indicate that our method obtains significant improvement over previous
 baselines.
****
Entropy-based adaptive Hamiltonian Monte Carlo
Marcel Hirt, Michalis Titsias, Petros Dellaportas
Hamiltonian Monte Carlo (HMC) is a popular Markov Chain Monte Carlo (MCMC) algor
ithm to sample from an unnormalized probability distribution. A leapfrog integra
tor is commonly used to implement HMC in practice, but its performance can be se
nsitive to the choice of mass matrix used therein. We develop a gradient-based a
lgorithm that allows for the adaptation of the mass matrix by encouraging the le
apfrog integrator to have high acceptance rates while also exploring all dimensi
ons jointly. In contrast to previous work that adapt the hyperparameters of HMC
using some form of expected squared jumping distance, the adaptation strategy su
ggested here aims to increase sampling efficiency by maximizing an approximation
 of the proposal entropy. We illustrate that using multiple gradients in the HMC
 proposal can be beneficial compared to a single gradient-step in Metropolis-adj
usted Langevin proposals. Empirical evidence suggests that the adaptation method
 can outperform different versions of HMC schemes by adjusting the mass matrix t
o the geometry of the target distribution and by providing some control on the i
ntegration time.
****
Continual World: A Robotic Benchmark For Continual Reinforcement Learning
Maciej Wo■czyk, Micha■ Zaj■c, Razvan Pascanu, ■ukasz Kuci■ski, Piotr Mi■o■
Continual learning (CL) --- the ability to continuously learn, building on previ
ously acquired knowledge --- is a natural requirement for long-lived autonomous
reinforcement learning (RL) agents. While building such agents, one needs to bal
ance opposing desiderata, such as constraints on capacity and compute, the abili
ty to not catastrophically forget, and to exhibit positive transfer on new tasks
. Understanding the right trade-off is conceptually and computationally challeng
ing, which we argue has led the community to overly focus on catastrophic forget
ting. In response to these issues, we advocate for the need to prioritize forwa
rd transfer and propose Continual World, a benchmark consisting of realistic and
 meaningfully diverse robotic tasks built on top of Meta-World as a testbed. Fo
llowing an in-depth empirical evaluation of existing CL methods, we pinpoint the
ir limitations and highlight unique algorithmic challenges in the RL setting. Ou
r benchmark aims to provide a meaningful and computationally inexpensive challen
ge for the community and thus help better understand the performance of existing

and future solutions. Information about the benchmark, including the open-source code, is available at https://sites.google.com/view/continualworld.
****

Towards Best-of-All-Worlds Online Learning with Feedback Graphs
Liad Erez, Tomer Koren
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

ViTAE: Vision Transformer Advanced by Exploring Intrinsic Inductive Bias
Yufei Xu, Qiming ZHANG, Jing Zhang, Dacheng Tao
Transformers have shown great potential in various computer vision tasks owing to their strong capability in modeling long-range dependency using the self-attention mechanism. Nevertheless, vision transformers treat an image as 1D sequence of visual tokens, lacking an intrinsic inductive bias (IB) in modeling local visual structures and dealing with scale variance. Alternatively, they require large-scale training data and longer training schedules to learn the IB implicitly. In this paper, we propose a new Vision Transformer Advanced by Exploring intrinsic IB from convolutions, i.e., ViTAE. Technically, ViTAE has several spatial pyramid reduction modules to downsample and embed the input image into tokens with rich multi-scale context by using multiple convolutions with different dilation rates. In this way, it acquires an intrinsic scale invariance IB and is able to learn robust feature representation for objects at various scales. Moreover, in each transformer layer, ViTAE has a convolution block in parallel to the multi-head self-attention module, whose features are fused and fed into the feed-forward network. Consequently, it has the intrinsic locality IB and is able to learn local features and global dependencies collaboratively. Experiments on ImageNet as well as downstream tasks prove the superiority of ViTAE over the baseline transformer and concurrent works. Source code and pretrained models will be available at https://github.com/Annbless/ViTAE.
****

Open Rule Induction
Wanyun Cui, Xingran Chen
Rules have a number of desirable properties. It is easy to understand, infer new knowledge, and communicate with other inference systems. One weakness of the previous rule induction systems is that they only find rules within a knowledge base (KB) and therefore cannot generalize to more open and complex real-world rules. Recently, the language model (LM)-based rule generation are proposed to enhance the expressive power of the rules.In this paper, we revisit the differences between KB-based rule induction and LM-based rule generation. We argue that, while KB-based methods inducted rules by discovering data commonalitiess, the current LM-based methods are learning rules from rules''. This limits these methods to only producecanned'' rules whose patterns are constrained by the annotated rules, while discarding the rich expressive power of LMs for free text.Therefore, in this paper, we propose the open rule induction problem, which aims to induce open rules utilizing the knowledge in LMs. Besides, we propose the Orion (\underline{o}pen \underline{r}ule \underline{i}nducti\underline{on}) system to automatically mine open rules from LMs without supervision of annotated rules. We conducted extensive experiments to verify the quality and quantity of the inducted open rules. Surprisingly, when applying the open rules in downstream tasks (i.e. relation extraction), these automatically inducted rules even outperformed the manually annotated rules.
****

Post-Contextual-Bandit Inference
Aurelien Bibaut, Maria Dimakopoulou, Nathan Kallus, Antoine Chambaz, Mark van der Laan
Contextual bandit algorithms are increasingly replacing non-adaptive A/B tests in e-commerce, healthcare, and policymaking because they can both improve outcomes for study participants and increase the chance of identifying good or even bes

t policies. To support credible inference on novel interventions at the end of t
he study, nonetheless, we still want to construct valid confidence intervals on
average treatment effects, subgroup effects, or value of new policies. The adapt
ive nature of the data collected by contextual bandit algorithms, however, makes
 this difficult: standard estimators are no longer asymptotically normally distr
ibuted and classic confidence intervals fail to provide correct coverage. While
this has been addressed in non-contextual settings by using stabilized estimator
s, variance stabilized estimators in the contextual setting pose unique challeng
es that we tackle for the first time in this paper. We propose the Contextual Ad
aptive Doubly Robust (CADR) estimator, a novel estimator for policy value that i
s asymptotically normal under contextual adaptive data collection. The main tech
nical challenge in constructing CADR is designing adaptive and consistent condit
ional standard deviation estimators for stabilization. Extensive numerical exper
iments using 57 OpenML datasets demonstrate that confidence intervals based on C
ADR uniquely provide correct coverage.
****
Revisiting Discriminator in GAN Compression: A Generator-discriminator Cooperati
ve Compression Scheme
Shaojie Li, Jie Wu, Xuefeng Xiao, Fei Chao, Xudong Mao, Rongrong Ji
Recently, a series of algorithms have been explored for GAN compression, which a
ims to reduce tremendous computational overhead and memory usages when deploying
 GANs on resource-constrained edge devices. However, most of the existing GAN co
mpression work only focuses on how to compress the generator, while fails to tak
e the discriminator into account. In this work, we revisit the role of discrimin
ator in GAN compression and design a novel generator-discriminator cooperative c
ompression scheme for GAN compression, termed GCC. Within GCC, a selective activ
ation discriminator automatically selects and activates convolutional channels a
ccording to a local capacity constraint and a global coordination constraint, wh
ich help maintain the Nash equilibrium with the lightweight generator during the
 adversarial training and avoid mode collapse. The original generator and discri
minator are also optimized from scratch, to play as a teacher model to progressi
vely refine the pruned generator and the selective activation discriminator. A n
ovel online collaborative distillation scheme is designed to take full advantage
 of the intermediate feature of the teacher generator and discriminator to furth
er boost the performance of the lightweight generator. Extensive experiments on
various GAN-based generation tasks demonstrate the effectiveness and generalizat
ion of GCC. Among them, GCC contributes to reducing 80% computational costs whil
e maintains comparable performance in image translation tasks.
****
Asymptotically Exact Error Characterization of Offline Policy Evaluation with Mi
sspecified Linear Models
Kohei Miyaguchi
We consider the problem of offline policy evaluation~(OPE) with Markov decision
processes~(MDPs), where the goal is to estimate the utility of given decision-ma
king policies based on static datasets. Recently, theoretical understanding of O
PE has been rapidly advanced under (approximate) realizability assumptions, i.e.
, where the environments of interest are well approximated with the given hypoth
etical models. On the other hand, the OPE under unrealizability has not been wel
l understood as much as in the realizable setting despite its importance in real
-world applications.To address this issue, we study the behavior of a simple exi
sting OPE method called the linear direct method~(DM) under the unrealizability.
 Consequently, we obtain an asymptotically exact characterization of the OPE err
or in a doubly robust form. Leveraging this result, we also establish the nonpar
ametric consistency of the tile-coding estimators under quite mild assumptions.
****
Topographic VAEs learn Equivariant Capsules
T. Anderson Keller, Max Welling
In this work we seek to bridge the concepts of topographic organization and equi
variance in neural networks. To accomplish this, we introduce the Topographic VA
E: a novel method for efficiently training deep generative models with topograph

ically organized latent variables. We show that such a model indeed learns to organize its activations according to salient characteristics such as digit class, width, and style on MNIST. Furthermore, through topographic organization over time (i.e. temporal coherence), we demonstrate how predefined latent space transformation operators can be encouraged for observed transformed input sequences -- a primitive form of unsupervised learned equivariance. We demonstrate that this model successfully learns sets of approximately equivariant features (i.e. "capsules") directly from sequences and achieves higher likelihood on correspondingly transforming test sequences. Equivariance is verified quantitatively by measuring the approximate commutativity of the inference network and the sequence transformations. Finally, we demonstrate approximate equivariance to complex transformations, expanding upon the capabilities of existing group equivariant neural networks.
****

MobILE: Model-Based Imitation Learning From Observation Alone
Rahul Kidambi, Jonathan Chang, Wen Sun
This paper studies Imitation Learning from Observations alone (ILFO) where the learner is presented with expert demonstrations that consist only of states visited by an expert (without access to actions taken by the expert). We present a provably efficient model-based framework MobILE to solve the ILFO problem. MobILE involves carefully trading off exploration against imitation - this is achieved by integrating the idea of optimism in the face of uncertainty into the distribution matching imitation learning (IL) framework. We provide a unified analysis for MobILE, and demonstrate that MobILE enjoys strong performance guarantees for classes of MDP dynamics that satisfy certain well studied notions of complexity. We also show that the ILFO problem is strictly harder than the standard IL problem by reducing ILFO to a multi-armed bandit problem indicating that exploration is necessary for solving ILFO efficiently. We complement these theoretical results with experimental simulations on benchmark OpenAI Gym tasks that indicate the efficacy of MobILE. Code for implementing the MobILE framework is available at https://github.com/rahulkidambi/MobILE-NeurIPS2021.
****

Few-Round Learning for Federated Learning
Younghyun Park, Dong-Jun Han, Do-Yeon Kim, Jun Seo, Jaekyun Moon
In federated learning (FL), a number of distributed clients targeting the same task collaborate to train a single global model without sharing their data. The learning process typically starts from a randomly initialized or some pretrained model. In this paper, we aim at designing an initial model based on which an arbitrary group of clients can obtain a global model for its own purpose, within only a few rounds of FL. The key challenge here is that the downstream tasks for which the pretrained model will be used are generally unknown when the initial model is prepared. Our idea is to take a meta-learning approach to construct the initial model so that any group with a possibly unseen task can obtain a high-accuracy global model within only R rounds of FL. Our meta-learning itself could be done via federated learning among willing participants and is based on an episodic arrangement to mimic the R rounds of FL followed by inference in each episode. Extensive experimental results show that our method generalizes well for arbitrary groups of clients and provides large performance improvements given the same overall communication/computation resources, compared to other baselines relying on known pretraining methods.
****

On Path Integration of Grid Cells: Group Representation and Isotropic Scaling
Ruiqi Gao, Jianwen Xie, Xue-Xin Wei, Song-Chun Zhu, Ying Nian Wu
Understanding how grid cells perform path integration calculations remains a fundamental problem. In this paper, we conduct theoretical analysis of a general representation model of path integration by grid cells, where the 2D self-position is encoded as a higher dimensional vector, and the 2D self-motion is represented by a general transformation of the vector. We identify two conditions on the transformation. One is a group representation condition that is necessary for path integration. The other is an isotropic scaling condition that ensures locally

conformal embedding, so that the error in the vector representation translates c onformally to the error in the 2D self-position. Then we investigate the simples t transformation, i.e., the linear transformation, uncover its explicit algebrai c and geometric structure as matrix Lie group of rotation, and explore the conne ction between the isotropic scaling condition and a special class of hexagon gri d patterns. Finally, with our optimization-based approach, we manage to learn he xagon grid patterns that share similar properties of the grid cells in the roden t brain. The learned model is capable of accurate long distance path integration . Code is available at https://github.com/ruiqigao/grid-cell-path.
****
Online Convex Optimization with Continuous Switching Constraint
Guanghui Wang, Yuanyu Wan, Tianbao Yang, Lijun 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.
****
Why Do Better Loss Functions Lead to Less Transferable Features?
Simon Kornblith, Ting Chen, Honglak Lee, Mohammad Norouzi
Previous work has proposed many new loss functions and regularizers that improve test accuracy on image classification tasks. However, it is not clear whether t hese loss functions learn better representations for downstream tasks. This pape r studies how the choice of training objective affects the transferability of th e hidden representations of convolutional neural networks trained on ImageNet. W e show that many objectives lead to statistically significant improvements in Im ageNet accuracy over vanilla softmax cross-entropy, but the resulting fixed feat ure extractors transfer substantially worse to downstream tasks, and the choice of loss has little effect when networks are fully fine-tuned on the new tasks. U sing centered kernel alignment to measure similarity between hidden representati ons of networks, we find that differences among loss functions are apparent only in the last few layers of the network. We delve deeper into representations of the penultimate layer, finding that different objectives and hyperparameter comb inations lead to dramatically different levels of class separation. Representati ons with higher class separation obtain higher accuracy on the original task, bu t their features are less useful for downstream tasks. Our results suggest there exists a trade-off between learning invariant features for the original task an d features relevant for transfer tasks.
****
Breaking the centralized barrier for cross-device federated learning
Sai Praneeth Karimireddy, Martin Jaggi, Satyen Kale, Mehryar Mohri, Sashank Redd i, Sebastian U. Stich, Ananda Theertha Suresh
Federated learning (FL) is a challenging setting for optimization due to the het erogeneity of the data across different clients which gives rise to the client d rift phenomenon. In fact, obtaining an algorithm for FL which is uniformly bette r than simple centralized training has been a major open problem thus far. In th is work, we propose a general algorithmic framework, Mime, which i) mitigates cl ient drift and ii) adapts arbitrary centralized optimization algorithms such as momentum and Adam to the cross-device federated learning setting. Mime uses a co mbination of control-variates and server-level statistics (e.g. momentum) at eve ry client-update step to ensure that each local update mimics that of the centra lized method run on iid data. We prove a reduction result showing that Mime can translate the convergence of a generic algorithm in the centralized setting into convergence in the federated setting. Further, we show that when combined with momentum based variance reduction, Mime is provably faster than any centralized method--the first such result. We also perform a thorough experimental explorati on of Mime's performance on real world datasets.
****
Adversarially robust learning for security-constrained optimal power flow
Priya Donti, Aayushya Agarwal, Neeraj Vijay Bedmutha, Larry Pileggi, J. Zico Kol ter

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 a Single Neuron with Bias Using Gradient Descent
Gal Vardi, Gilad Yehudai, Ohad Shamir
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Making a (Counterfactual) Difference One Rationale at a Time
Mitchell Plyler, Michael Green, Min Chi
Rationales, snippets of extracted text that explain an inference, have emerged as a popular framework for interpretable natural language processing (NLP). Rationale models typically consist of two cooperating modules: a selector and a classifier with the goal of maximizing the mutual information (MMI) between the "selected" text and the document label. Despite their promises, MMI-based methods often pick up on spurious text patterns and result in models with nonsensical behaviors. In this work, we investigate whether counterfactual data augmentation (CDA), without human assistance, can improve the performance of the selector by lowering the mutual information between spurious signals and the document label. Our counterfactuals are produced in an unsupervised fashion using class-dependent generative models. From an information theoretic lens, we derive properties of the unaugmented dataset for which our CDA approach would succeed. The effectiveness of CDA is empirically evaluated by comparing against several baselines including an improved MMI-based rationale schema on two multi-aspect datasets. Our results show that CDA produces rationales that better capture the signal of interest.
****
3D Siamese Voxel-to-BEV Tracker for Sparse Point Clouds
Le Hui, Lingpeng Wang, Mingmei Cheng, Jin Xie, Jian Yang
3D object tracking in point clouds is still a challenging problem due to the sparsity of LiDAR points in dynamic environments. In this work, we propose a Siamese voxel-to-BEV tracker, which can significantly improve the tracking performance in sparse 3D point clouds. Specifically, it consists of a Siamese shape-aware feature learning network and a voxel-to-BEV target localization network. The Siamese shape-aware feature learning network can capture 3D shape information of the object to learn the discriminative features of the object so that the potential target from the background in sparse point clouds can be identified. To this end, we first perform template feature embedding to embed the template's feature into the potential target and then generate a dense 3D shape to characterize the shape information of the potential target. For localizing the tracked target, the voxel-to-BEV target localization network regresses the target's 2D center and the z-axis center from the dense bird's eye view (BEV) feature map in an anchor-free manner. Concretely, we compress the voxelized point cloud along z-axis through max pooling to obtain a dense BEV feature map, where the regression of the 2D center and the z-axis center can be performed more effectively. Extensive evaluation on the KITTI tracking dataset shows that our method significantly outperforms the current state-of-the-art methods by a large margin. Code is available at https://github.com/fpthink/V2B.
****
Stateful Strategic Regression
Keegan Harris, Hoda Heidari, Steven Z. Wu
Automated decision-making tools increasingly assess individuals to determine if they qualify for high-stakes opportunities. A recent line of research investigates how strategic agents may respond to such scoring tools to receive favorable assessments. While prior work has focused on the short-term strategic interactions between a decision-making institution (modeled as a principal) and individual

decision-subjects (modeled as agents), we investigate interactions spanning multiple time-steps. In particular, we consider settings in which the agent's effort investment today can accumulate over time in the form of an internal state - impacting both his future rewards and that of the principal. We characterize the Stackelberg equilibrium of the resulting game and provide novel algorithms for computing it. Our analysis reveals several intriguing insights about the role of multiple interactions in shaping the game's outcome: First, we establish that in our stateful setting, the class of all linear assessment policies remains as powerful as the larger class of all monotonic assessment policies. While recovering the principal's optimal policy requires solving a non-convex optimization problem, we provide polynomial-time algorithms for recovering both the principal and agent's optimal policies under common assumptions about the process by which effort investments convert to observable features. Most importantly, we show that with multiple rounds of interaction at her disposal, the principal is more effective at incentivizing the agent to accumulate effort in her desired direction. Our work addresses several critical gaps in the growing literature on the societal impacts of automated decision-making - by focusing on longer time horizons and accounting for the compounding nature of decisions individuals receive over time.
****

Self-Attention Between Datapoints: Going Beyond Individual Input-Output Pairs in Deep Learning

Jannik Kossen, Neil Band, Clare Lyle, Aidan N. Gomez, Thomas Rainforth, Yarin Gal

We challenge a common assumption underlying most supervised deep learning: that a model makes a prediction depending only on its parameters and the features of a single input. To this end, we introduce a general-purpose deep learning architecture that takes as input the entire dataset instead of processing one datapoint at a time. Our approach uses self-attention to reason about relationships between datapoints explicitly, which can be seen as realizing non-parametric models using parametric attention mechanisms. However, unlike conventional non-parametric models, we let the model learn end-to-end from the data how to make use of other datapoints for prediction. Empirically, our models solve cross-datapoint lookup and complex reasoning tasks unsolvable by traditional deep learning models. We show highly competitive results on tabular data, early results on CIFAR-10, and give insight into how the model makes use of the interactions between points.
****

Your head is there to move you around: Goal-driven models of the primate dorsal pathway

Patrick Mineault, Shahab Bakhtiari, Blake Richards, Christopher Pack

Neurons in the dorsal visual pathway of the mammalian brain are selective for motion stimuli, with the complexity of stimulus representations increasing along the hierarchy. This progression is similar to that of the ventral visual pathway, which is well characterized by artificial neural networks (ANNs) optimized for object recognition. In contrast, there are no image-computable models of the dorsal stream with comparable explanatory power. We hypothesized that the properties of dorsal stream neurons could be explained by a simple learning objective: the need for an organism to orient itself during self-motion. To test this hypothesis, we trained a 3D ResNet to predict an agent's self-motion parameters from visual stimuli in a simulated environment. We found that the responses in this network accounted well for the selectivity of neurons in a large database of single-neuron recordings from the dorsal visual stream of non-human primates. In contrast, ANNs trained on an action recognition dataset through supervised or self-supervised learning could not explain responses in the dorsal stream, despite also being trained on naturalistic videos with moving objects. These results demonstrate that an ecologically relevant cost function can account for dorsal stream properties in the primate brain.
****

Achieving Rotational Invariance with Bessel-Convolutional Neural Networks

Valentin Delchevalerie, Adrien Bibal, Benoît Frénay, Alexandre Mayer

For many applications in image analysis, learning models that are invariant to translations and rotations is paramount. This is the case, for example, in medical imaging where the objects of interest can appear at arbitrary positions, with arbitrary orientations. As of today, Convolutional Neural Networks (CNN) are one of the most powerful tools for image analysis. They achieve, thanks to convolutions, an invariance with respect to translations. In this work, we present a new type of convolutional layer that takes advantage of Bessel functions, well known in physics, to build Bessel-CNNs (B-CNNs) that are invariant to all the continuous set of possible rotation angles by design.
****

Unsupervised Domain Adaptation with Dynamics-Aware Rewards in Reinforcement Learning

Jinxin Liu, Hao Shen, Donglin Wang, Yachen Kang, Qiangxing Tian

Unsupervised reinforcement learning aims to acquire skills without prior goal representations, where an agent automatically explores an open-ended environment to represent goals and learn the goal-conditioned policy. However, this procedure is often time-consuming, limiting the rollout in some potentially expensive target environments. The intuitive approach of training in another interaction-rich environment disrupts the reproducibility of trained skills in the target environment due to the dynamics shifts and thus inhibits direct transferring. Assuming free access to a source environment, we propose an unsupervised domain adaptation method to identify and acquire skills across dynamics. Particularly, we introduce a KL regularized objective to encourage emergence of skills, rewarding the agent for both discovering skills and aligning its behaviors respecting dynamics shifts. This suggests that both dynamics (source and target) shape the reward to facilitate the learning of adaptive skills. We also conduct empirical experiments to demonstrate that our method can effectively learn skills that can be smoothly deployed in target.
****

GraphFormers: GNN-nested Transformers for Representation Learning on Textual Graph

Junhan Yang, Zheng Liu, Shitao Xiao, Chaozhuo Li, Defu Lian, Sanjay Agrawal, Amit Singh, Guangzhong Sun, Xing Xie

The representation learning on textual graph is to generate low-dimensional embeddings for the nodes based on the individual textual features and the neighbourhood information. Recent breakthroughs on pretrained language models and graph neural networks push forward the development of corresponding techniques. The existing works mainly rely on the cascaded model architecture: the textual features of nodes are independently encoded by language models at first; the textual embeddings are aggregated by graph neural networks afterwards. However, the above architecture is limited due to the independent modeling of textual features. In this work, we propose GraphFormers, where layerwise GNN components are nested alongside the transformer blocks of language models. With the proposed architecture, the text encoding and the graph aggregation are fused into an iterative workflow, making each node's semantic accurately comprehended from the global perspective. In addition, a progressive learning strategy is introduced, where the model is successively trained on manipulated data and original data to reinforce its capability of integrating information on graph. Extensive evaluations are conducted on three large-scale benchmark datasets, where GraphFormers outperform the SOTA baselines with comparable running efficiency. The source code is released at https://github.com/microsoft/GraphFormers .
****

A Universal Law of Robustness via Isoperimetry

Sebastien Bubeck, Mark Sellke

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

Emile Mathieu, Adam Foster, Yee Teh
Learning representations of stochastic processes is an emerging problem in machi
ne learning with applications from meta-learning to physical object models to ti
me series. Typical methods rely on exact reconstruction of observations, but thi
s approach breaks down as observations become high-dimensional or noise distribu
tions become complex. To address this, we propose a unifying framework for learn
ing contrastive representations of stochastic processes (CReSP) that does away w
ith exact reconstruction. We dissect potential use cases for stochastic process
representations, and propose methods that accommodate each. Empirically, we show
that our methods are effective for learning representations of periodic functio
ns, 3D objects and dynamical processes. Our methods tolerate noisy high-dimensio
nal observations better than traditional approaches, and the learned representat
ions transfer to a range of downstream tasks.
****
A Domain-Shrinking based Bayesian Optimization Algorithm with Order-Optimal Regr
et Performance
Sudeep Salgia, Sattar Vakili, Qing Zhao
Requests for name changes in the electronic proceedings will be accepted with no
questions asked. However name changes may cause bibliographic tracking issues.
Authors are asked to consider this carefully and discuss it with their co-auth
ors prior to requesting a name change in the electronic proceedings.
****
Scalars are universal: Equivariant machine learning, structured like classical p
hysics
Soledad Villar, David W Hogg, Kate Storey-Fisher, Weichi Yao, Ben Blum-Smith
Requests for name changes in the electronic proceedings will be accepted with 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.
****
Unsupervised Object-Level Representation Learning from Scene Images
Jiahao Xie, Xiaohang Zhan, Ziwei Liu, Yew Soon Ong, Chen Change Loy
Contrastive self-supervised learning has largely narrowed the gap to supervised
pre-training on ImageNet. However, its success highly relies on the object-centr
ic priors of ImageNet, i.e., different augmented views of the same image corresp
ond to the same object. Such a heavily curated constraint becomes immediately in
feasible when pre-trained on more complex scene images with many objects. To ove
rcome this limitation, we introduce Object-level Representation Learning (ORL),
a new self-supervised learning framework towards scene images. Our key insight i
s to leverage image-level self-supervised pre-training as the prior to discover
object-level semantic correspondence, thus realizing object-level representation
learning from scene images. Extensive experiments on COCO show that ORL signifi
cantly improves the performance of self-supervised learning on scene images, eve
n surpassing supervised ImageNet pre-training on several downstream tasks. Furth
ermore, ORL improves the downstream performance when more unlabeled scene images
are available, demonstrating its great potential of harnessing unlabeled data i
n the wild. We hope our approach can motivate future research on more general-pu
rpose unsupervised representation learning from scene data.
****
Do Transformers Really Perform Badly for Graph Representation?
Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanmin
g Shen, Tie-Yan Liu
The Transformer architecture has become a dominant choice in many domains, such
as natural language processing and computer vision. Yet, it has not achieved com
petitive performance on popular leaderboards of graph-level prediction compared
to mainstream GNN variants. Therefore, it remains a mystery how Transformers cou
ld perform well for graph representation learning. In this paper, we solve this
mystery by presenting Graphormer, which is built upon the standard Transformer a
rchitecture, and could attain excellent results on a broad range of graph repres
entation learning tasks, especially on the recent OGB Large-Scale Challenge. Our

key insight to utilizing Transformer in the graph is the necessity of effective
ly encoding the structural information of a graph into the model. To this end, w
e propose several simple yet effective structural encoding methods to help Graph
ormer better model graph-structured data. Besides, we mathematically characteriz
e the expressive power of Graphormer and exhibit that with our ways of encoding
the structural information of graphs, many popular GNN variants could be covered
 as the special cases of Graphormer. The code and models of Graphormer will be m
ade publicly available at \url{https://github.com/Microsoft/Graphormer}.
****
Powerpropagation: A sparsity inducing weight reparameterisation
Jonathan Schwarz, Siddhant Jayakumar, Razvan Pascanu, Peter E Latham, Yee Teh
The training of sparse neural networks is becoming an increasingly important too
l for reducing the computational footprint of models at training and evaluation,
 as well enabling the effective scaling up of models. Whereas much work over the
 years has been dedicated to specialised pruning techniques, little attention ha
s been paid to the inherent effect of gradient based training on model sparsity.
 Inthis work, we introduce Powerpropagation, a new weight-parameterisation for n
eural networks that leads to inherently sparse models. Exploiting the behaviour
of gradient descent, our method gives rise to weight updates exhibiting a “rich
get richer” dynamic, leaving low-magnitude parameters largely unaffected by lear
ning. Models trained in this manner exhibit similar performance, but have a dist
ributionwith markedly higher density at zero, allowing more parameters to be pru
ned safely. Powerpropagation is general, intuitive, cheap and straight-forward t
o implement and can readily be combined with various other techniques. To highli
ght its versatility, we explore it in two very different settings: Firstly, foll
owing a recent line of work, we investigate its effect on sparse training for re
source-constrained settings. Here, we combine Powerpropagation with a traditiona
l weight-pruning technique as well as recent state-of-the-art sparse-to-sparse a
lgorithms, showing superior performance on the ImageNet benchmark. Secondly, we
advocate the useof sparsity in overcoming catastrophic forgetting, where compres
sed representations allow accommodating a large number of tasks at fixed model c
apacity. In all cases our reparameterisation considerably increases the efficacy
 of the off-the-shelf methods.
****
Stronger NAS with Weaker Predictors
Junru Wu, Xiyang Dai, Dongdong Chen, Yinpeng Chen, Mengchen Liu, Ye Yu, Zhangyan
g Wang, Zicheng Liu, Mei Chen, Lu Yuan
Neural Architecture Search (NAS) often trains and evaluates a large number of ar
chitectures. Recent predictor-based NAS approaches attempt to alleviate such hea
vy computation costs with two key steps: sampling some architecture-performance
pairs and fitting a proxy accuracy predictor. Given limited samples, these predi
ctors, however, are far from accurate to locate top architectures due to the dif
ficulty of fitting the huge search space. This paper reflects on a simple yet cr
ucial question: if our final goal is to find the best architecture, do we really
 need to model the whole space well?. We propose a paradigm shift from fitting t
he whole architecture space using one strong predictor, to progressively fitting
 a search path towards the high-performance sub-space through a set of weaker pr
edictors. As a key property of the weak predictors, their probabilities of sampl
ing better architectures keep increasing. Hence we only sample a few well-perfor
med architectures guided by the previously learned predictor and estimate a new
better weak predictor. This embarrassingly easy framework, dubbed WeakNAS, produ
ces coarse-to-fine iteration to gradually refine the ranking of sampling space.
Extensive experiments demonstrate that WeakNAS costs fewer samples to find top-p
erformance architectures on NAS-Bench-101 and NAS-Bench-201. Compared to state-o
f-the-art (SOTA) predictor-based NAS methods, WeakNAS outperforms all with notab
le margins, e.g., requiring at least 7.5x less samples to find global optimal on
 NAS-Bench-101. WeakNAS can also absorb their ideas to boost performance more. F
urther, WeakNAS strikes the new SOTA result of 81.3% in the ImageNet MobileNet S
earch Space. The code is available at: https://github.com/VITA-Group/WeakNAS.
****

Convolutional Normalization: Improving Deep Convolutional Network Robustness and Training
Sheng Liu, Xiao Li, Yuexiang Zhai, Chong You, Zhihui Zhu, Carlos Fernandez-Granda, Qing Qu
Normalization techniques have become a basic component in modern convolutional neural networks (ConvNets). In particular, many recent works demonstrate that promoting the orthogonality of the weights helps train deep models and improve robustness. For ConvNets, most existing methods are based on penalizing or normalizing weight matrices derived from concatenating or flattening the convolutional kernels. These methods often destroy or ignore the benign convolutional structure of the kernels; therefore, they are often expensive or impractical for deep ConvNets. In contrast, we introduce a simple and efficient ``Convolutional Normalization'' (ConvNorm) method that can fully exploit the convolutional structure in the Fourier domain and serve as a simple plug-and-play module to be conveniently incorporated into any ConvNets. Our method is inspired by recent work on preconditioning methods for convolutional sparse coding and can effectively promote each layer's channel-wise isometry. Furthermore, we show that our ConvNorm can reduce the layerwise spectral norm of the weight matrices and hence improve the Lipschitzness of the network, leading to easier training and improved robustness for deep ConvNets. Applied to classification under noise corruptions and generative adversarial network (GAN), we show that the ConvNorm improves the robustness of common ConvNets such as ResNet and the performance of GAN. We verify our findings via numerical experiments on CIFAR and ImageNet. Our implementation is available online at \url{https://github.com/shengliu66/ConvNorm}.
****
Nearly-Tight and Oblivious Algorithms for Explainable Clustering
Buddhima Gamlath, Xinrui Jia, Adam Polak, 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-authors prior to requesting a name change in the electronic proceedings.
****
Deep Networks Provably Classify Data on Curves
Tingran Wang, Sam Buchanan, Dar Gilboa, John Wright
Data with low-dimensional nonlinear structure are ubiquitous in engineering and scientific problems. We study a model problem with such structure---a binary classification task that uses a deep fully-connected neural network to classify data drawn from two disjoint smooth curves on the unit sphere. Aside from mild regularity conditions, we place no restrictions on the configuration of the curves. We prove that when (i) the network depth is large relative to certain geometric properties that set the difficulty of the problem and (ii) the network width and number of samples is polynomial in the depth, randomly-initialized gradient descent quickly learns to correctly classify all points on the two curves with high probability. To our knowledge, this is the first generalization guarantee for deep networks with nonlinear data that depends only on intrinsic data properties. Our analysis proceeds by a reduction to dynamics in the neural tangent kernel (NTK) regime, where the network depth plays the role of a fitting resource in solving the classification problem. In particular, via fine-grained control of the decay properties of the NTK, we demonstrate that when the network is sufficiently deep, the NTK can be locally approximated by a translationally invariant operator on the manifolds and stably inverted over smooth functions, which guarantees convergence and generalization.
****
COMBO: Conservative Offline Model-Based Policy Optimization
Tianhe Yu, Aviral Kumar, Rafael Rafailov, Aravind Rajeswaran, Sergey Levine, Chelsea Finn
Model-based reinforcement learning (RL) algorithms, which learn a dynamics model from logged experience and perform conservative planning under the learned model, have emerged as a promising paradigm for offline reinforcement learning (offline RL). However, practical variants of such model-based algorithms rely on expl

icit uncertainty quantification for incorporating conservatism. Uncertainty esti mation with complex models, such as deep neural networks, can be difficult and u nreliable. We empirically find that uncertainty estimation is not accurate and l eads to poor performance in certain scenarios in offline model-based RL. We over come this limitation by developing a new model-based offline RL algorithm, COMBO , that trains a value function using both the offline dataset and data generated using rollouts under the model while also additionally regularizing the value f unction on out-of-support state-action tuples generated via model rollouts. This results in a conservative estimate of the value function for out-of-support sta te-action tuples, without requiring explicit uncertainty estimation. Theoretical ly, we show that COMBO satisfies a policy improvement guarantee in the offline s etting. Through extensive experiments, we find that COMBO attains greater perfor mance compared to prior offline RL on problems that demand generalization to rel ated but previously unseen tasks, and also consistently matches or outperforms p rior offline RL methods on widely studied offline RL benchmarks, including image -based tasks.
****

Time-series Generation by Contrastive Imitation

Daniel Jarrett, Ioana Bica, Mihaela van der Schaar

Consider learning a generative model for time-series data. The sequential settin g poses a unique challenge: Not only should the generator capture the conditiona l dynamics of (stepwise) transitions, but its open-loop rollouts should also pre serve the joint distribution of (multi-step) trajectories. On one hand, autoregr essive models trained by MLE allow learning and computing explicit transition di stributions, but suffer from compounding error during rollouts. On the other han d, adversarial models based on GAN training alleviate such exposure bias, but tr ansitions are implicit and hard to assess. In this work, we study a generative f ramework that seeks to combine the strengths of both: Motivated by a moment-matc hing objective to mitigate compounding error, we optimize a local (but forward-l ooking) transition policy, where the reinforcement signal is provided by a globa l (but stepwise-decomposable) energy model trained by contrastive estimation. At training, the two components are learned cooperatively, avoiding the instabilit ies typical of adversarial objectives. At inference, the learned policy serves a s the generator for iterative sampling, and the learned energy serves as a traje ctory-level measure for evaluating sample quality. By expressly training a polic y to imitate sequential behavior of time-series features in a dataset, this appr oach embodies "generation by imitation". Theoretically, we illustrate the correc tness of this formulation and the consistency of the algorithm. Empirically, we evaluate its ability to generate predictively useful samples from real-world dat asets, verifying that it performs at the standard of existing benchmarks.
****

Differentially Private Sampling from Distributions

Sofya Raskhodnikova, Satchit Sivakumar, Adam Smith, Marika Swanberg

Requests for name changes in the electronic proceedings will be accepted with 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 the Expected Complexity of Maxout Networks

Hanna Tseran, Guido F. Montufar

Learning with neural networks relies on the complexity of their representable fu nctions, but more importantly, their particular assignment of typical parameters to functions of different complexity. Taking the number of activation regions a s a complexity measure, recent works have shown that the practical complexity of deep ReLU networks is often far from the theoretical maximum. In this work, we show that this phenomenon also occurs in networks with maxout (multi-argument) a ctivation functions and when considering the decision boundaries in classificati on tasks. We also show that the parameter space has a multitude of full-dimensio nal regions with widely different complexity, and obtain nontrivial lower bounds on the expected complexity. Finally, we investigate different parameter initial

ization procedures and show that they can increase the speed of convergence in training.
****
Cross-view Geo-localization with Layer-to-Layer Transformer
Hongji Yang, Xiufan Lu, Yingying Zhu
In this work, we address the problem of cross-view geo-localization, which estimates the geospatial location of a street view image by matching it with a database of geo-tagged aerial images. The cross-view matching task is extremely challenging due to drastic appearance and geometry differences across views. Unlike existing methods that predominantly fall back on CNN, here we devise a novel layer-to-layer Transformer (L2LTR) that utilizes the properties of self-attention in Transformer to model global dependencies, thus significantly decreasing visual ambiguities in cross-view geo-localization. We also exploit the positional encoding of the Transformer to help the L2LTR understand and correspond geometric configurations between ground and aerial images. Compared to state-of-the-art methods that impose strong assumptions on geometry knowledge, the L2LTR flexibly learns the positional embeddings through the training objective. It hence becomes more practical in many real-world scenarios. Although Transformer is well suited to our task, its vanilla self-attention mechanism independently interacts within image patches in each layer, which overlooks correlations between layers. Instead, this paper proposes a simple yet effective self-cross attention mechanism to improve the quality of learned representations. Self-cross attention models global dependencies between adjacent layers and creates short paths for effective information flow. As a result, the proposed self-cross attention leads to more stable training, improves the generalization ability, and prevents the learned intermediate features from being overly similar. Extensive experiments demonstrate that our L2LTR performs favorably against state-of-the-art methods on standard, fine-grained, and cross-dataset cross-view geo-localization tasks. The code is available online.
****
TAAC: Temporally Abstract Actor-Critic for Continuous Control
Haonan Yu, Wei Xu, Haichao Zhang
We present temporally abstract actor-critic (TAAC), a simple but effective off-policy RL algorithm that incorporates closed-loop temporal abstraction into the actor-critic framework. TAAC adds a second-stage binary policy to choose between the previous action and a new action output by an actor. Crucially, its "act-or-repeat" decision hinges on the actually sampled action instead of the expected behavior of the actor. This post-acting switching scheme let the overall policy make more informed decisions. TAAC has two important features: a) persistent exploration, and b) a new compare-through Q operator for multi-step TD backup, specially tailored to the action repetition scenario. We demonstrate TAAC's advantages over several strong baselines across 14 continuous control tasks. Our surprising finding reveals that while achieving top performance, TAAC is able to "mine" a significant number of repeated actions with the trained policy even on continuous tasks whose problem structures on the surface seem to repel action repetition. This suggests that aside from encouraging persistent exploration, action repetition can find its place in a good policy behavior. Code is available at https://github.com/hnyu/taac.
****
Learning Robust Hierarchical Patterns of Human Brain across Many fMRI Studies
Dushyant Sahoo, Christos Davatzikos
Multi-site fMRI studies face the challenge that the pooling introduces systematic non-biological site-specific variance due to hardware, software, and environment. In this paper, we propose to reduce site-specific variance in the estimation of hierarchical Sparsity Connectivity Patterns (hSCPs) in fMRI data via a simple yet effective matrix factorization while preserving biologically relevant variations. Our method leverages unsupervised adversarial learning to improve the reproducibility of the components. Experiments on simulated datasets display that the proposed method can estimate components with higher accuracy and reproducibility, while preserving age-related variation on a multi-center clinical data set

.
****

Global Convergence to Local Minmax Equilibrium in Classes of Nonconvex Zero-Sum Games

Tanner Fiez, Lillian Ratliff, Eric Mazumdar, Evan Faulkner, Adhyyan Narang

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

Aditya Gopalan, Braghadeesh Lakshminarayanan, Venkatesh Saligrama

Many industrial and security applications employ a suite of sensors for detecting abrupt changes in temporal behavior patterns. These abrupt changes typically manifest locally, rendering only a small subset of sensors informative. Continuous monitoring of every sensor can be expensive due to resource constraints, and serves as a motivation for the bandit quickest changepoint detection problem, where sensing actions (or sensors) are sequentially chosen, and only measurements corresponding to chosen actions are observed. We derive an information-theoretic lower bound on the detection delay for a general class of finitely parameterized probability distributions. We then propose a computationally efficient online sensing scheme, which seamlessly balances the need for exploration of different sensing options with exploitation of querying informative actions. We derive expected delay bounds for the proposed scheme and show that these bounds match our information-theoretic lower bounds at low false alarm rates, establishing optimality of the proposed method. We then perform a number of experiments on synthetic and real datasets demonstrating the effectiveness of our proposed method.

****

Can multi-label classification networks know what they don’t know?

Haoran Wang, Weitang Liu, Alex Bocchieri, Yixuan Li

Estimating out-of-distribution (OOD) uncertainty is a major challenge for safely deploying machine learning models in the open-world environment. Improved methods for OOD detection in multi-class classification have emerged, while OOD detection methods for multi-label classification remain underexplored and use rudimentary techniques. We propose JointEnergy, a simple and effective method, which estimates the OOD indicator scores by aggregating label-wise energy scores from multiple labels. We show that JointEnergy can be mathematically interpreted from a joint likelihood perspective. Our results show consistent improvement over previous methods that are based on the maximum-valued scores, which fail to capture joint information from multiple labels. We demonstrate the effectiveness of our method on three common multi-label classification benchmarks, including MS-COCO, PASCAL-VOC, and NUS-WIDE. We show that JointEnergy can reduce the FPR95 by up to 10.05% compared to the previous best baseline, establishing state-of-the-art performance.

****

Balanced Chamfer Distance as a Comprehensive Metric for Point Cloud Completion

Tong Wu, Liang Pan, Junzhe Zhang, Tai WANG, Ziwei Liu, Dahua Lin

Chamfer Distance (CD) and Earth Mover’s Distance (EMD) are two broadly adopted metrics for measuring the similarity between two point sets. However, CD is usually insensitive to mismatched local density, and EMD is usually dominated by global distribution while overlooks the fidelity of detailed structures. Besides, their unbounded value range induces a heavy influence from the outliers. These defects prevent them from providing a consistent evaluation. To tackle these problems, we propose a new similarity measure named Density-aware Chamfer Distance (DCD). It is derived from CD and benefits from several desirable properties: 1) it can detect disparity of density distributions and is thus a more intensive measure of similarity compared to CD; 2) it is stricter with detailed structures and significantly more computationally efficient than EMD; 3) the bounded value range encourages a more stable and reasonable evaluation over the whole test set. We adopt DCD to evaluate the point cloud completion task, where experimental resul

ts show that DCD pays attention to both the overall structure and local geometric details and provides a more reliable evaluation even when CD and EMD contradict each other. We can also use DCD as the training loss, which outperforms the same model trained with CD loss on all three metrics. In addition, we propose a novel point discriminator module that estimates the priority for another guided down-sampling step, and it achieves noticeable improvements under DCD together with competitive results for both CD and EMD. We hope our work could pave the way for a more comprehensive and practical point cloud similarity evaluation. Our code will be available at https://github.com/wutong16/DensityawareChamfer_Distance.
****

Optimal Gradient-based Algorithms for Non-concave Bandit Optimization
Baihe Huang, Kaixuan Huang, Sham Kakade, Jason D. Lee, Qi Lei, Runzhe Wang, Jiaqi Yang
Bandit problems with linear or concave reward have been extensively studied, but relatively few works have studied bandits with non-concave reward. This work considers a large family of bandit problems where the unknown underlying reward function is non-concave, including the low-rank generalized linear bandit problems and two-layer neural network with polynomial activation bandit problem.For the low-rank generalized linear bandit problem, we provide a minimax-optimal algorithm in the dimension, refuting both conjectures in \cite{lu2021low,jun2019bilinear}. Our algorithms are based on a unified zeroth-order optimization paradigm that applies in great generality and attains optimal rates in several structured polynomial settings (in the dimension). We further demonstrate the applicability of our algorithms in RL in the generative model setting, resulting in improved sample complexity over prior approaches.Finally, we show that the standard optimistic algorithms (e.g., UCB) are sub-optimal by dimension factors. In the neural net setting (with polynomial activation functions) with noiseless reward, we provide a bandit algorithm with sample complexity equal to the intrinsic algebraic dimension. Again, we show that optimistic approaches have worse sample complexity, polynomial in the extrinsic dimension (which could be exponentially worse in the polynomial degree).
****

On Optimal Interpolation in Linear Regression
Eduard Oravkin, Patrick Rebeschini
Understanding when and why interpolating methods generalize well has recently been a topic of interest in statistical learning theory. However, systematically connecting interpolating methods to achievable notions of optimality has only received partial attention. In this paper, we ask the question of what is the optimal way to interpolate in linear regression using functions that are linear in the response variable (as the case for the Bayes optimal estimator in ridge regression) and depend on the data, the population covariance of the data, the signal-to-noise ratio and the covariance of the prior for the signal, but do not depend on the value of the signal itself nor the noise vector in the training data. We provide a closed-form expression for the interpolator that achieves this notion of optimality and show that it can be derived as the limit of preconditioned gradient descent with a specific initialization. We identify a regime where the minimum-norm interpolator provably generalizes arbitrarily worse than the optimal response-linear achievable interpolator that we introduce, and validate with numerical experiments that the notion of optimality we consider can be achieved by interpolating methods that only use the training data as input in the case of an isotropic prior. Finally, we extend the notion of optimal response-linear interpolation to random features regression under a linear data-generating model.
****

Differentiable Optimization of Generalized Nondecomposable Functions using Linear Programs
Zihang Meng, Lopamudra Mukherjee, Yichao Wu, Vikas Singh, Sathya Ravi
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****
Towards Understanding Cooperative Multi-Agent Q-Learning with Value Factorizatio
n
Jianhao Wang, Zhizhou Ren, Beining Han, Jianing Ye, Chongjie Zhang
Value factorization is a popular and promising approach to scaling up multi-agen
t reinforcement learning in cooperative settings, which balances the learning sc
alability and the representational capacity of value functions. However, the the
oretical understanding of such methods is limited. In this paper, we formalize a
 multi-agent fitted Q-iteration framework for analyzing factorized multi-agent Q
-learning. Based on this framework, we investigate linear value factorization an
d reveal that multi-agent Q-learning with this simple decomposition implicitly r
ealizes a powerful counterfactual credit assignment, but may not converge in som
e settings. Through further analysis, we find that on-policy training or richer
joint value function classes can improve its local or global convergence propert
ies, respectively. Finally, to support our theoretical implications in practical
 realization, we conduct an empirical analysis of state-of-the-art deep multi-ag
ent Q-learning algorithms on didactic examples and a broad set of StarCraft II u
nit micromanagement tasks.
****
Margin-Independent Online Multiclass Learning via Convex Geometry
Guru Guruganesh, Allen Liu, Jon Schneider, Joshua 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.
****
STEP: Out-of-Distribution Detection in the Presence of Limited In-Distribution L
abeled Data
Zhi Zhou, Lan-Zhe Guo, Zhanzhan Cheng, Yu-Feng Li, Shiliang Pu
Existing semi-supervised learning (SSL) studies typically assume that unlabeled
and test data are drawn from the same distribution as labeled data. However, in
many real-world applications, it is desirable to have SSL algorithms that not on
ly classify the samples drawn from the same distribution of labeled data but als
o detect out-of-distribution (OOD) samples drawn from an unknown distribution. I
n this paper, we study a setting called semi-supervised OOD detection. Two main
challenges compared with previous OOD detection settings are i) the lack of labe
led data and in-distribution data; ii) OOD samples could be unseen during traini
ng. Efforts on this direction remain limited. In this paper, we present an appro
ach STEP significantly improving OOD detection performance by introducing a new
technique: Structure-Keep Unzipping. It learns a new representation space in whi
ch OOD samples could be separated well. An efficient optimization algorithm is d
erived to solve the objective. Comprehensive experiments across various OOD dete
ction benchmarks clearly show that our STEP approach outperforms other methods b
y a large margin and achieves remarkable detection performance on several benchm
arks.
****
Renyi Differential Privacy of The Subsampled Shuffle Model In Distributed Learni
ng
Antonious Girgis, Deepesh Data, Suhas Diggavi
We study privacy in a distributed learning framework, where clients collaborativ
ely build a learning model iteratively throughinteractions with a server from wh
om we need privacy. Motivated by stochastic optimization and the federated learn
ing (FL) paradigm, we focus on the case where a small fraction of data samples a
re randomly sub-sampled in each round to participate in the learning process, wh
ich also enables privacy amplification. To obtain even stronger local privacy g
uarantees, we study this in the shuffle privacy model, where each client randomi
zes its response using a local differentially private (LDP) mechanism and the se
rver only receives a random permutation (shuffle) of the clients' responses with
out theirassociation to each client. The principal result of this paper is a pri
vacy-optimization performance trade-off for discrete randomization mechanisms in

this sub-sampled shuffle privacy model. This is enabledthrough a new theoretical technique to analyze the Renyi Differential Privacy (RDP) of the sub-sampled shuffle model. We numerically demonstrate that, for important regimes, with composition our boundyields significant improvement in privacy guarantee over the state-of-the-art approximate Differential Privacy (DP) guarantee (with strong composition) for sub-sampled shuffled models. We also demonstrate numerically significant improvement in privacy-learning performance operating point using real data sets. Despite these advances, an open question is to bridge the gap between lower and upper privacy bounds in our RDP analysis.
****

Gradient-based Editing of Memory Examples for Online Task-free Continual Learning

Xisen Jin, Arka Sadhu, Junyi Du, Xiang Ren

We explore task-free continual learning (CL), in which a model is trained to avoid catastrophic forgetting in the absence of explicit task boundaries or identities. Among many efforts on task-free CL, a notable family of approaches are memory-based that store and replay a subset of training examples. However, the utility of stored seen examples may diminish over time since CL models are continually updated. Here, we propose Gradient based Memory EDiting (GMED), a framework for editing stored examples in continuous input space via gradient updates, in order to create more "challenging" examples for replay. GMED-edited examples remain similar to their unedited forms, but can yield increased loss in the upcoming model updates, thereby making the future replays more effective in overcoming catastrophic forgetting. By construction, GMED can be seamlessly applied in conjunction with other memory-based CL algorithms to bring further improvement. Experiments validate the effectiveness of GMED, and our best method significantly outperforms baselines and previous state-of-the-art on five out of six datasets.
****

Tailoring: encoding inductive biases by optimizing unsupervised objectives at prediction time

Ferran Alet, Maria Bauza, Kenji Kawaguchi, Nurullah Giray Kuru, Tomás Lozano-Pérez, Leslie Kaelbling

From CNNs to attention mechanisms, encoding inductive biases into neural networks has been a fruitful source of improvement in machine learning. Adding auxiliary losses to the main objective function is a general way of encoding biases that can help networks learn better representations. However, since auxiliary losses are minimized only on training data, they suffer from the same generalization gap as regular task losses. Moreover, by adding a term to the loss function, the model optimizes a different objective than the one we care about. In this work we address both problems: first, we take inspiration from transductive learning and note that after receiving an input but before making a prediction, we can fine-tune our networks on any unsupervised loss. We call this process tailoring, because we customize the model to each input to ensure our prediction satisfies the inductive bias. Second, we formulate meta-tailoring, a nested optimization similar to that in meta-learning, and train our models to perform well on the task objective after adapting them using an unsupervised loss. The advantages of tailoring and meta-tailoring are discussed theoretically and demonstrated empirically on a diverse set of examples.
****

Implicit Bias of SGD for Diagonal Linear Networks: a Provable Benefit of Stochasticity

Scott Pesme, Loucas Pillaud-Vivien, Nicolas Flammarion

Understanding the implicit bias of training algorithms is of crucial importance in order to explain the success of overparametrised neural networks. In this paper, we study the dynamics of stochastic gradient descent over diagonal linear networks through its continuous time version, namely stochastic gradient flow. We explicitly characterise the solution chosen by the stochastic flow and prove that it always enjoys better generalisation properties than that of gradient flow.Quite surprisingly, we show that the convergence speed of the training loss controls the magnitude of the biasing effect: the slower the convergence, the better

the bias. To fully complete our analysis, we provide convergence guarantees for the dynamics. We also give experimental results which support our theoretical claims. Our findings highlight the fact that structured noise can induce better generalisation and they help explain the greater performances of stochastic gradient descent over gradient descent observed in practice.
****

Iterative Teacher-Aware Learning
Luyao Yuan, Dongruo Zhou, Junhong Shen, Jingdong Gao, Jeffrey L Chen, Quanquan Gu, Ying Nian Wu, Song-Chun Zhu
In human pedagogy, teachers and students can interact adaptively to maximize communication efficiency. The teacher adjusts her teaching method for different students, and the student, after getting familiar with the teacher's instruction mechanism, can infer the teacher's intention to learn faster. Recently, the benefits of integrating this cooperative pedagogy into machine concept learning in discrete spaces have been proved by multiple works. However, how cooperative pedagogy can facilitate machine parameter learning hasn't been thoroughly studied. In this paper, we propose a gradient optimization based teacher-aware learner who can incorporate teacher's cooperative intention into the likelihood function and learn provably faster compared with the naive learning algorithms used in previous machine teaching works. We give theoretical proof that the iterative teacher-aware learning (ITAL) process leads to local and global improvements. We then validate our algorithms with extensive experiments on various tasks including regression, classification, and inverse reinforcement learning using synthetic and real data. We also show the advantage of modeling teacher-awareness when agents are learning from human teachers.
****

Clockwork Variational Autoencoders
Vaibhav Saxena, Jimmy Ba, Danijar Hafner
Deep learning has enabled algorithms to generate realistic images. However, accurately predicting long video sequences requires understanding long-term dependencies and remains an open challenge. While existing video prediction models succeed at generating sharp images, they tend to fail at accurately predicting far into the future. We introduce the Clockwork VAE (CW-VAE), a video prediction model that leverages a hierarchy of latent sequences, where higher levels tick at slower intervals. We demonstrate the benefits of both hierarchical latents and temporal abstraction on 4 diverse video prediction datasets with sequences of up to 1000 frames, where CW-VAE outperforms top video prediction models. Additionally, we propose a Minecraft benchmark for long-term video prediction. We conduct several experiments to gain insights into CW-VAE and confirm that slower levels learn to represent objects that change more slowly in the video, and faster levels learn to represent faster objects.
****

How Does it Sound?
Kun Su, Xiulong Liu, Eli Shlizerman
One of the primary purposes of video is to capture people and their unique activities. It is often the case that the experience of watching the video can be enhanced by adding a musical soundtrack that is in-sync with the rhythmic features of these activities. How would this soundtrack sound? Such a problem is challenging since little is known about capturing the rhythmic nature of free body movements. In this work, we explore this problem and propose a novel system, called `RhythmicNet', which takes as an input a video which includes human movements and generates a soundtrack for it. RhythmicNet works directly with human movements by extracting skeleton keypoints and implements a sequence of models which translate the keypoints to rhythmic sounds.RhythmicNet follows the natural process of music improvisation which includes the prescription of streams of the beat, the rhythm and the melody. In particular, RhythmicNet first infers the music beat and the style pattern from body keypoints per each frame to produce rhythm. Next, it implements a transformer-based model to generate the hits of drum instruments and implements a U-net based model to generate the velocity and the offsets of the instruments. Additional types of instruments are added to the soundtrack by

further conditioning on the generated drum sounds. We evaluate RhythmicNet on large scale datasets of videos that include body movements with inherit sound association, such as dance, as well as 'in the wild' internet videos of various movements and actions. We show that the method can generate plausible music that aligns well with different types of human movements.
****

Stabilizing Dynamical Systems via Policy Gradient Methods
Juan Perdomo, Jack Umenberger, Max Simchowitz
Stabilizing an unknown control system is one of the most fundamental problems in control systems engineering. In this paper, we provide a simple, model-free algorithm for stabilizing fully observed dynamical systems. While model-free methods have become increasingly popular in practice due to their simplicity and flexibility, stabilization via direct policy search has received surprisingly little attention. Our algorithm proceeds by solving a series of discounted LQR problems, where the discount factor is gradually increased. We prove that this method efficiently recovers a stabilizing controller for linear systems, and for smooth, nonlinear systems within a neighborhood of their equilibria. Our approach overcomes a significant limitation of prior work, namely the need for a pre-given stabilizing control policy. We empirically evaluate the effectiveness of our approach on common control benchmarks.
****

Language models enable zero-shot prediction of the effects of mutations on protein function
Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, Alex Rives
Modeling the effect of sequence variation on function is a fundamental problem for understanding and designing proteins. Since evolution encodes information about function into patterns in protein sequences, unsupervised models of variant effects can be learned from sequence data. The approach to date has been to fit a model to a family of related sequences. The conventional setting is limited, since a new model must be trained for each prediction task. We show that using only zero-shot inference, without any supervision from experimental data or additional training, protein language models capture the functional effects of sequence variation, performing at state-of-the-art.
****

Deep Reinforcement Learning at the Edge of the Statistical Precipice
Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C. Courville, Marc Bellemare
Deep reinforcement learning (RL) algorithms are predominantly evaluated by comparing their relative performance on a large suite of tasks. Most published results on deep RL benchmarks compare point estimates of aggregate performance such as mean and median scores across tasks, ignoring the statistical uncertainty implied by the use of a finite number of training runs. Beginning with the Arcade Learning Environment (ALE), the shift towards computationally-demanding benchmarks has led to the practice of evaluating only a small number of runs per task, exacerbating the statistical uncertainty in point estimates. In this paper, we argue that reliable evaluation in the few run deep RL regime cannot ignore the uncertainty in results without running the risk of slowing down progress in the field. We illustrate this point using a case study on the Atari 100k benchmark, where we find substantial discrepancies between conclusions drawn from point estimates alone versus a more thorough statistical analysis. With the aim of increasing the field's confidence in reported results with a handful of runs, we advocate for reporting interval estimates of aggregate performance and propose performance profiles to account for the variability in results, as well as present more robust and efficient aggregate metrics, such as interquartile mean scores, to achieve small uncertainty in results. Using such statistical tools, we scrutinize performance evaluations of existing algorithms on other widely used RL benchmarks including the ALE, Procgen, and the DeepMind Control Suite, again revealing discrepancies in prior comparisons. Our findings call for a change in how we evaluate performance in deep RL, for which we present a more rigorous evaluation methodology, accompanied with an open-source library rliable, to prevent unreliable resu

lts from stagnating the field. This work received an outstanding paper award at NeurIPS 2021.
****
DRONE: Data-aware Low-rank Compression for Large NLP Models
Patrick Chen, Hsiang-Fu Yu, Inderjit Dhillon, Cho-Jui Hsieh
The representations learned by large-scale NLP models such as BERT have been widely used in various tasks. However, the increasing model size of the pre-trained models also brings efficiency challenges, including inference speed and model size when deploying models on mobile devices. Specifically, most operations in BERT consist of matrix multiplications. These matrices are not low-rank and thus canonical matrix decomposition could not find an efficient approximation. In this paper, we observe that the learned representation of each layer lies in a low-dimensional space. Based on this observation, we propose DRONE (data-aware low-rank compression), a provably optimal low-rank decomposition of weight matrices, which has a simple closed form solution that can be efficiently computed. DRONE can be applied to both fully connected and self-attention layers appearing in the BERT model. In addition to compressing standard models, out method can also be used on distilled BERT models to further improve compression rate. Experimental results show that DRONE is able to improve both model size and inference speed with limited loss in accuracy. Specifically, DRONE alone achieves 1.92x speedup on the MRPC task with only 1.5% loss in accuracy, and when DRONE is combined with distillation, it further achieves over 12.3x speedup on various natural language inference tasks.
****
DSelect-k: Differentiable Selection in the Mixture of Experts with Applications to Multi-Task Learning
Hussein Hazimeh, Zhe Zhao, Aakanksha Chowdhery, Maheswaran Sathiamoorthy, Yihua Chen, Rahul Mazumder, Lichan Hong, Ed Chi
The Mixture-of-Experts (MoE) architecture is showing promising results in improving parameter sharing in multi-task learning (MTL) and in scaling high-capacity neural networks. State-of-the-art MoE models use a trainable "sparse gate'" to select a subset of the experts for each input example. While conceptually appealing, existing sparse gates, such as Top-k, are not smooth. The lack of smoothness can lead to convergence and statistical performance issues when training with gradient-based methods. In this paper, we develop DSelect-k: a continuously differentiable and sparse gate for MoE, based on a novel binary encoding formulation. The gate can be trained using first-order methods, such as stochastic gradient descent, and offers explicit control over the number of experts to select. We demonstrate the effectiveness of DSelect-k on both synthetic and real MTL datasets with up to 128 tasks. Our experiments indicate that DSelect-k can achieve statistically significant improvements in prediction and expert selection over popular MoE gates. Notably, on a real-world, large-scale recommender system, DSelect-k achieves over 22% improvement in predictive performance compared to Top-k. We provide an open-source implementation of DSelect-k.
****
Mind the Gap: Assessing Temporal Generalization in Neural Language Models
Angeliki Lazaridou, Adhi Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomas Kocisky, Sebastian Ruder, Dani Yogatama, Kris Cao, Susannah Young, Phil Blunsom
Our world is open-ended, non-stationary, and constantly evolving; thus what we talk about and how we talk about it change over time. This inherent dynamic nature of language contrasts with the current static language modelling paradigm, which trains and evaluates models on utterances from overlapping time periods. Despite impressive recent progress, we demonstrate that Transformer-XL language models perform worse in the realistic setup of predicting future utterances from beyond their training period, and that model performance becomes increasingly worse with time. We find that, while increasing model size alone—a key driver behind recent progress—does not solve this problem, having models that continually update their knowledge with new information can indeed mitigate this performance degradation over time. Hence, given the compilation of ever-larger language modelli

ng datasets, combined with the growing list of language-model-based NLP applications that require up-to-date factual knowledge about the world, we argue that now is the right time to rethink the static way in which we currently train and evaluate our language models, and develop adaptive language models that can remain up-to-date with respect to our ever-changing and non-stationary world. We publicly release our dynamic, streaming language modelling benchmarks for WMT and arXiv to facilitate language model evaluation that takes temporal dynamics into account.
****

Heavy Tails in SGD and Compressibility of Overparametrized Neural Networks
Melih Barsbey, Milad Sefidgaran, Murat A. Erdogdu, Gaël Richard, 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-authors prior to requesting a name change in the electronic proceedings.
****

Targeted Neural Dynamical Modeling
Cole Hurwitz, Akash Srivastava, Kai Xu, Justin Jude, Matthew Perich, Lee Miller, Matthias Hennig
Latent dynamics models have emerged as powerful tools for modeling and interpreting neural population activity. Recently, there has been a focus on incorporating simultaneously measured behaviour into these models to further disentangle sources of neural variability in their latent space. These approaches, however, are limited in their ability to capture the underlying neural dynamics (e.g. linear) and in their ability to relate the learned dynamics back to the observed behaviour (e.g. no time lag). To this end, we introduce Targeted Neural Dynamical Modeling (TNDM), a nonlinear state-space model that jointly models the neural activity and external behavioural variables. TNDM decomposes neural dynamics into behaviourally relevant and behaviourally irrelevant dynamics; the relevant dynamics are used to reconstruct the behaviour through a flexible linear decoder and both sets of dynamics are used to reconstruct the neural activity through a linear decoder with no time lag. We implement TNDM as a sequential variational autoencoder and validate it on simulated recordings and recordings taken from the premotor and motor cortex of a monkey performing a center-out reaching task. We show that TNDM is able to learn low-dimensional latent dynamics that are highly predictive of behaviour without sacrificing its fit to the neural data.
****

Exploiting the Intrinsic Neighborhood Structure for Source-free Domain Adaptation
Shiqi Yang, yaxing wang, Joost van de Weijer, Luis Herranz, Shangling Jui
Domain adaptation (DA) aims to alleviate the domain shift between source domain and target domain. Most DA methods require access to the source data, but often that is not possible (e.g. due to data privacy or intellectual property). In this paper, we address the challenging source-free domain adaptation (SFDA) problem, where the source pretrained model is adapted to the target domain in the absence of source data. Our method is based on the observation that target data, which might no longer align with the source domain classifier, still forms clear clusters. We capture this intrinsic structure by defining local affinity of the target data, and encourage label consistency among data with high local affinity. We observe that higher affinity should be assigned to reciprocal neighbors, and propose a self regularization loss to decrease the negative impact of noisy neighbors. Furthermore, to aggregate information with more context, we consider expanded neighborhoods with small affinity values. In the experimental results we verify that the inherent structure of the target features is an important source of information for domain adaptation. We demonstrate that this local structure can be efficiently captured by considering the local neighbors, the reciprocal neighbors, and the expanded neighborhood. Finally, we achieve state-of-the-art performance on several 2D image and 3D point cloud recognition datasets. Code is available in https://github.com/Albert0147/SFDA_neighbors.
****

Learning with Noisy Correspondence for Cross-modal Matching
Zhenyu Huang, Guocheng Niu, Xiao Liu, Wenbiao Ding, Xinyan Xiao, Hua Wu, Xi Peng
Cross-modal matching, which aims to establish the correspondence between two different modalities, is fundamental to a variety of tasks such as cross-modal retrieval and vision-and-language understanding. Although a huge number of cross-modal matching methods have been proposed and achieved remarkable progress in recent years, almost all of these methods implicitly assume that the multimodal training data are correctly aligned. In practice, however, such an assumption is extremely expensive even impossible to satisfy. Based on this observation, we reveal and study a latent and challenging direction in cross-modal matching, named noisy correspondence, which could be regarded as a new paradigm of noisy labels. Different from the traditional noisy labels which mainly refer to the errors in category labels, our noisy correspondence refers to the mismatch paired samples. To solve this new problem, we propose a novel method for learning with noisy correspondence, named Noisy Correspondence Rectifier (NCR). In brief, NCR divides the data into clean and noisy partitions based on the memorization effect of neural networks and then rectifies the correspondence via an adaptive prediction model in a co-teaching manner. To verify the effectiveness of our method, we conduct experiments by using the image-text matching as a showcase. Extensive experiments on Flickr30K, MS-COCO, and Conceptual Captions verify the effectiveness of our method. The code could be accessed from www.pengxi.me .
****
Offline Reinforcement Learning with Reverse Model-based Imagination
Jianhao Wang, Wenzhe Li, Haozhe Jiang, Guangxiang Zhu, Siyuan Li, Chongjie Zhang
In offline reinforcement learning (offline RL), one of the main challenges is to deal with the distributional shift between the learning policy and the given dataset. To address this problem, recent offline RL methods attempt to introduce conservatism bias to encourage learning in high-confidence areas. Model-free approaches directly encode such bias into policy or value function learning using conservative regularizations or special network structures, but their constrained policy search limits the generalization beyond the offline dataset. Model-based approaches learn forward dynamics models with conservatism quantifications and then generate imaginary trajectories to extend the offline datasets. However, due to limited samples in offline datasets, conservatism quantifications often suffer from overgeneralization in out-of-support regions. The unreliable conservative measures will mislead forward model-based imaginations to undesired areas, leading to overaggressive behaviors. To encourage more conservatism, we propose a novel model-based offline RL framework, called Reverse Offline Model-based Imagination (ROMI). We learn a reverse dynamics model in conjunction with a novel reverse policy, which can generate rollouts leading to the target goal states within the offline dataset. These reverse imaginations provide informed data augmentation for model-free policy learning and enable conservative generalization beyond the offline dataset. ROMI can effectively combine with off-the-shelf model-free algorithms to enable model-based generalization with proper conservatism. Empirical results show that our method can generate more conservative behaviors and achieve state-of-the-art performance on offline RL benchmark tasks.
****
Parameter Prediction for Unseen Deep Architectures
Boris Knyazev, Michal Drozdzal, Graham W. Taylor, Adriana Romero Soriano
Deep learning has been successful in automating the design of features in machine learning pipelines. However, the algorithms optimizing neural network parameters remain largely hand-designed and computationally inefficient. We study if we can use deep learning to directly predict these parameters by exploiting the past knowledge of training other networks. We introduce a large-scale dataset of diverse computational graphs of neural architectures - DeepNets-1M - and use it to explore parameter prediction on CIFAR-10 and ImageNet. By leveraging advances in graph neural networks, we propose a hypernetwork that can predict performant parameters in a single forward pass taking a fraction of a second, even on a CPU. The proposed model achieves surprisingly good performance on unseen and diverse networks. For example, it is able to predict all 24 million parameters of a Res

Net-50 achieving a 60% accuracy on CIFAR-10. On ImageNet, top-5 accuracy of some of our networks approaches 50%. Our task along with the model and results can potentially lead to a new, more computationally efficient paradigm of training networks. Our model also learns a strong representation of neural architectures enabling their analysis.
****
FMMformer: Efficient and Flexible Transformer via Decomposed Near-field and Far-field Attention
Tan Nguyen, Vai Suliafu, Stanley Osher, Long Chen, Bao 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.
****
Square Root Principal Component Pursuit: Tuning-Free Noisy Robust Matrix Recovery
Junhui Zhang, Jingkai Yan, John Wright
We propose a new framework -- Square Root Principal Component Pursuit -- for low-rank matrix recovery from observations corrupted with noise and outliers. Inspired by the square root Lasso, this new formulation does not require prior knowledge of the noise level. We show that a single, universal choice of the regularization parameter suffices to achieve reconstruction error proportional to the (a priori unknown) noise level. In comparison, previous formulations such as stable PCP rely on noise-dependent parameters to achieve similar performance, and are therefore challenging to deploy in applications where the noise level is unknown. We validate the effectiveness of our new method through experiments on simulated and real datasets. Our simulations corroborate the claim that a universal choice of the regularization parameter yields near optimal performance across a range of noise levels, indicating that the proposed method outperforms the (somewhat loose) bound proved here.
****
Neural Bellman-Ford Networks: A General Graph Neural Network Framework for Link Prediction
Zhaocheng Zhu, Zuobai Zhang, Louis-Pascal Xhonneux, Jian Tang
Link prediction is a very fundamental task on graphs. Inspired by traditional path-based methods, in this paper we propose a general and flexible representation learning framework based on paths for link prediction. Specifically, we define the representation of a pair of nodes as the generalized sum of all path representations, with each path representation as the generalized product of the edge representations in the path. Motivated by the Bellman-Ford algorithm for solving the shortest path problem, we show that the proposed path formulation can be efficiently solved by the generalized Bellman-Ford algorithm. To further improve the capacity of the path formulation, we propose the Neural Bellman-Ford Network (NBFNet), a general graph neural network framework that solves the path formulation with learned operators in the generalized Bellman-Ford algorithm. The NBFNet parameterizes the generalized Bellman-Ford algorithm with 3 neural components, namely Indicator, Message and Aggregate functions, which corresponds to the boundary condition, multiplication operator, and summation operator respectively. The NBFNet covers many traditional path-based methods, and can be applied to both homogeneous graphs and multi-relational graphs (e.g., knowledge graphs) in both transductive and inductive settings. Experiments on both homogeneous graphs and knowledge graphs show that the proposed NBFNet outperforms existing methods by a large margin in both transductive and inductive settings, achieving new state-of-the-art results.
****
CorticalFlow: A Diffeomorphic Mesh Transformer Network for Cortical Surface Reconstruction
Leo Lebrat, Rodrigo Santa Cruz, Frederic de Gournay, Darren Fu, Pierrick Bourgeat, Jurgen Fripp, Clinton Fookes, Olivier Salvado
In this paper, we introduce CorticalFlow, a new geometric deep-learning model th

at, given a 3-dimensional image, learns to deform a reference template towards a targeted object. To conserve the template mesh’s topological properties, we train our model over a set of diffeomorphic transformations. This new implementation of a flow Ordinary Differential Equation (ODE) framework benefits from a small GPU memory footprint, allowing the generation of surfaces with several hundred thousand vertices. To reduce topological errors introduced by its discrete resolution, we derive numeric conditions which improve the manifoldness of the predicted triangle mesh. To exhibit the utility of CorticalFlow, we demonstrate its performance for the challenging task of brain cortical surface reconstruction. In contrast to the current state-of-the-art, CorticalFlow produces superior surfaces while reducing the computation time from nine and a half minutes to one second. More significantly, CorticalFlow enforces the generation of anatomically plausible surfaces; the absence of which has been a major impediment restricting the clinical relevance of such surface reconstruction methods.
****
Bridging the Gap Between Practice and PAC-Bayes Theory in Few-Shot Meta-Learning
Nan Ding, Xi Chen, Tomer Levinboim, Sebastian Goodman, Radu Soricut
Despite recent advances in its theoretical understanding, there still remains a significant gap in the ability of existing PAC-Bayesian theories on meta-learning to explain performance improvements in the few-shot learning setting, where the number of training examples in the target tasks is severely limited. This gap originates from an assumption in the existing theories which supposes that the number of training examples in the observed tasks and the number of training examples in the target tasks follow the same distribution, an assumption that rarely holds in practice. By relaxing this assumption, we develop two PAC-Bayesian bounds tailored for the few-shot learning setting and show that two existing meta-learning algorithms (MAML and Reptile) can be derived from our bounds, thereby bridging the gap between practice and PAC-Bayesian theories. Furthermore, we derive a new computationally-efficient PACMAML algorithm, and show it outperforms existing meta-learning algorithms on several few-shot benchmark datasets.
****
SLOE: A Faster Method for Statistical Inference in High-Dimensional Logistic Regression
Steve Yadlowsky, Taedong Yun, Cory Y McLean, Alexander D'Amour
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
ELLA: Exploration through Learned Language Abstraction
Suvir Mirchandani, Siddharth Karamcheti, Dorsa Sadigh
Building agents capable of understanding language instructions is critical to effective and robust human-AI collaboration. Recent work focuses on training these agents via reinforcement learning in environments with synthetic language; however, instructions often define long-horizon, sparse-reward tasks, and learning policies requires many episodes of experience. We introduce ELLA: Exploration through Learned Language Abstraction, a reward shaping approach geared towards boosting sample efficiency in sparse reward environments by correlating high-level instructions with simpler low-level constituents. ELLA has two key elements: 1) A termination classifier that identifies when agents complete low-level instructions, and 2) A relevance classifier that correlates low-level instructions with success on high-level tasks. We learn the termination classifier offline from pairs of instructions and terminal states. Notably, in departure from prior work in language and abstraction, we learn the relevance classifier online, without relying on an explicit decomposition of high-level instructions to low-level instructions. On a suite of complex BabyAI environments with varying instruction complexities and reward sparsity, ELLA shows gains in sample efficiency relative to language-based shaping and traditional RL methods.
****
Learning Distilled Collaboration Graph for Multi-Agent Perception

Yiming Li, Shunli Ren, Pengxiang Wu, Siheng Chen, Chen Feng, Wenjun Zhang

To promote better performance-bandwidth trade-off for multi-agent perception, we propose a novel distilled collaboration graph (DiscoGraph) to model trainable, pose-aware, and adaptive collaboration among agents. Our key novelties lie in two aspects. First, we propose a teacher-student framework to train DiscoGraph via knowledge distillation. The teacher model employs an early collaboration with holistic-view inputs; the student model is based on intermediate collaboration with single-view inputs. Our framework trains DiscoGraph by constraining post-collaboration feature maps in the student model to match the correspondences in the teacher model. Second, we propose a matrix-valued edge weight in DiscoGraph. In such a matrix, each element reflects the inter-agent attention at a specific spatial region, allowing an agent to adaptively highlight the informative regions. During inference, we only need to use the student model named as the distilled collaboration network (DiscoNet). Attributed to the teacher-student framework, multiple agents with the shared DiscoNet could collaboratively approach the performance of a hypothetical teacher model with a holistic view. Our approach is validated on V2X-Sim 1.0, a large-scale multi-agent perception dataset that we synthesized using CARLA and SUMO co-simulation. Our quantitative and qualitative experiments in multi-agent 3D object detection show that DiscoNet could not only achieve a better performance-bandwidth trade-off than the state-of-the-art collaborative perception methods, but also bring more straightforward design rationale. Our code is available on https://github.com/ai4ce/DiscoNet.

****

Federated-EM with heterogeneity mitigation and variance reduction

Aymeric Dieuleveut, Gersende Fort, Eric Moulines, Geneviève Robin

The Expectation Maximization (EM) algorithm is the default algorithm for inference in latent variable models. As in any other field of machine learning, applications of latent variable models to very large datasets make the use of advanced parallel and distributed architecture mandatory. This paper introduces FedEM, which is the first extension of the EM algorithm to the federated learning context. FedEM is a new communication efficient method, which handles partial participation of local devices, and is robust to heterogeneous distribution of the datasets. To alleviate the communication bottleneck, FedEM compresses appropriately defined complete data sufficient statistics. We also develop and analyze an extension of FedEM to further incorporate a variance reduction scheme. In all cases, we derive finite-time complexity bounds for smooth non-convex problems. Numerical results are presented to support our theoretical findings, as well as an application to federated missing values imputation for biodiversity monitoring.

****

On the Role of Optimization in Double Descent: A Least Squares Study

Ilja Kuzborskij, Csaba Szepesvari, Omar Rivasplata, Amal Rannen-Triki, Razvan Pascanu

Empirically it has been observed that the performance of deep neural networks steadily improves with increased model size, contradicting the classical view on overfitting and generalization. Recently, the double descent phenomenon has been proposed to reconcile this observation with theory, suggesting that the test error has a second descent when the model becomes sufficiently overparameterized, as the model size itself acts as an implicit regularizer. In this paper we add to the growing body of work in this space, providing a careful study of learning dynamics as a function of model size for the least squares scenario. We show an excess risk bound for the gradient descent solution of the least squares objective. The bound depends on the smallest non-zero eigenvalue of the sample covariance matrix of the input features, via a functional form that has the double descent behaviour. This gives a new perspective on the double descent curves reported in the literature, as our analysis of the excess risk allows to decouple the effect of optimization and generalization error. In particular, we find that in the case of noiseless regression, double descent is explained solely by optimization-related quantities, which was missed in studies focusing on the Moore-Penrose pseudoinverse solution. We believe that our derivation provides an alternative view compared to existing works, shedding some light on a possible cause of this

phenomenon, at least in the considered least squares setting. We empirically exp lore if our predictions hold for neural networks, in particular whether the spec trum of the sample covariance of features at intermediary hidden layers has a si milar behaviour as the one predicted by our derivations in the least squares set ting.
****
Neural Architecture Dilation for Adversarial Robustness
Yanxi Li, Zhaohui Yang, Yunhe Wang, Chang Xu
With the tremendous advances in the architecture and scale of convolutional neur al networks (CNNs) over the past few decades, they can easily reach or even exce ed the performance of humans in certain tasks. However, a recently discovered sh ortcoming of CNNs is that they are vulnerable to adversarial attacks. Although t he adversarial robustness of CNNs can be improved by adversarial training, there is a trade-off between standard accuracy and adversarial robustness. From the n eural architecture perspective, this paper aims to improve the adversarial robus tness of the backbone CNNs that have a satisfactory accuracy. Under a minimal co mputational overhead, the introduction of a dilation architecture is expected to be friendly with the standard performance of the backbone CNN while pursuing ad versarial robustness. Theoretical analyses on the standard and adversarial error bounds naturally motivate the proposed neural architecture dilation algorithm. Experimental results on real-world datasets and benchmark neural networks demons trate the effectiveness of the proposed algorithm to balance the accuracy and ad versarial robustness.
****
Clustering Effect of Adversarial Robust Models
Yang Bai, Xin Yan, Yong Jiang, Shu-Tao Xia, Yisen Wang
Adversarial robustness has received increasing attention along with the study of adversarial examples. So far, existing works show that robust models not only o btain robustness against various adversarial attacks but also boost the performa nce in some downstream tasks. However, the underlying mechanism of adversarial r obustness is still not clear. In this paper, we interpret adversarial robustness from the perspective of linear components, and find that there exist some stati stical properties for comprehensively robust models. Specifically, robust models show obvious hierarchical clustering effect on their linearized sub-networks, w hen removing or replacing all non-linear components (e.g., batch normalization, maximum pooling, or activation layers). Based on these observations, we propose a novel understanding of adversarial robustness and apply it on more tasks inclu ding domain adaption and robustness boosting. Experimental evaluations demonstra te the rationality and superiority of our proposed clustering strategy. Our code is available at https://github.com/bymavis/AdvWeightNeurIPS2021.
****
On the Cryptographic Hardness of Learning Single Periodic Neurons
Min Jae Song, Ilias Zadik, Joan Bruna
Requests for name changes in the electronic proceedings will be accepted with 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.
****
PCA Initialization for Approximate Message Passing in Rotationally Invariant Mod els
Marco Mondelli, Ramji Venkataramanan
We study the problem of estimating a rank-1 signal in the presence of rotational ly invariant noise--a class of perturbations more general than Gaussian noise. Principal Component Analysis (PCA) provides a natural estimator, and sharp resul ts on its performance have been obtained in the high-dimensional regime. Recentl y, an Approximate Message Passing (AMP) algorithm has been proposed as an altern ative estimator with the potential to improve the accuracy of PCA. However, the existing analysis of AMP requires an initialization that is both correlated with the signal and independent of the noise, which is often unrealistic in practice . In this work, we combine the two methods, and propose to initialize AMP with P

CA. Our main result is a rigorous asymptotic characterization of the performance of this estimator. Both the AMP algorithm and its analysis differ from those previously derived in the Gaussian setting: at every iteration, our AMP algorithm requires a specific term to account for PCA initialization, while in the Gaussian case, PCA initialization affects only the first iteration of AMP. The proof is based on a two-phase artificial AMP that first approximates the PCA estimator and then mimics the true AMP. Our numerical simulations show an excellent agreement between AMP results and theoretical predictions, and suggest an interesting open direction on achieving Bayes-optimal performance.
****
Automatic and Harmless Regularization with Constrained and Lexicographic Optimization: A Dynamic Barrier Approach
Chengyue Gong, Xingchao Liu, Qiang Liu
Many machine learning tasks have to make a trade-off between two loss functions, typically the main data-fitness loss and an auxiliary loss. The most widely used approach is to optimize the linear combination of the objectives, which, however, requires manual tuning of the combination coefficient and is theoretically unsuitable for non-convex functions. In this work, we consider constrained optimization as a more principled approach for trading off two losses, with a special emphasis on lexicographic optimization, a degenerated limit of constrained optimization which optimizes a secondary loss inside the optimal set of the main loss. We propose a dynamic barrier gradient descent algorithm which provides a unified solution of both constrained and lexicographic optimization. We establish the convergence of the method for general non-convex functions.
****
Corruption Robust Active Learning
Yifang Chen, Simon S. Du, Kevin G. Jamieson
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Metadata-based Multi-Task Bandits with Bayesian Hierarchical Models
Runzhe Wan, Lin Ge, Rui Song
How to explore efficiently is a central problem in multi-armed bandits. In this paper, we introduce the metadata-based multi-task bandit problem, where the agent needs to solve a large number of related multi-armed bandit tasks and can leverage some task-specific features (i.e., metadata) to share knowledge across tasks. As a general framework, we propose to capture task relations through the lens of Bayesian hierarchical models, upon which a Thompson sampling algorithm is designed to efficiently learn task relations, share information, and minimize the cumulative regrets. Two concrete examples for Gaussian bandits and Bernoulli bandits are carefully analyzed. The Bayes regret for Gaussian bandits clearly demonstrates the benefits of information sharing with our algorithm. The proposed method is further supported by extensive experiments.
****
Program Synthesis Guided Reinforcement Learning for Partially Observed Environments
Yichen Yang, Jeevana Priya Inala, Osbert Bastani, Yewen Pu, Armando Solar-Lezama, Martin Rinard
A key challenge for reinforcement learning is solving long-horizon planning problems. Recent work has leveraged programs to guide reinforcement learning in these settings. However, these approaches impose a high manual burden on the user since they must provide a guiding program for every new task. Partially observed environments further complicate the programming task because the program must implement a strategy that correctly, and ideally optimally, handles every possible configuration of the hidden regions of the environment. We propose a new approach, model predictive program synthesis (MPPS), that uses program synthesis to automatically generate the guiding programs. It trains a generative model to predict the unobserved portions of the world, and then synthesizes a program based on

samples from this model in a way that is robust to its uncertainty. In our exper
iments, we show that our approach significantly outperforms non-program-guided a
pproaches on a set of challenging benchmarks, including a 2D Minecraft-inspired
environment where the agent must complete a complex sequence of subtasks to achi
eve its goal, and achieves a similar performance as using handcrafted programs t
o guide the agent. Our results demonstrate that our approach can obtain the bene
fits of program-guided reinforcement learning without requiring the user to prov
ide a new guiding program for every new task.
****
Robust Allocations with Diversity Constraints
Zeyu Shen, Lodewijk Gelauff, Ashish Goel, Aleksandra Korolova, Kamesh Munagala
We consider the problem of allocating divisible items among multiple agents, and
consider the setting where any agent is allowed to introduce {\emph diversity c
onstraints} on the items they are allocated. We motivate this via settings where
the items themselves correspond to user ad slots or task workers with attribute
s such as race and gender on which the principal seeks to achieve demographic pa
rity. We consider the following question: When an agent expresses diversity cons
traints into an allocation rule, is the allocation of other agents hurt signific
antly? If this happens, the cost of introducing such constraints is disproportio
nately borne by agents who do not benefit from diversity. We codify this via two
desiderata capturing {\em robustness}. These are {\emph no negative externality
} -- other agents are not hurt -- and {\emph monotonicity} -- the agent enforcin
g the constraint does not see a large increase in value. We show in a formal sen
se that the Nash Welfare rule that maximizes product of agent values is {\emph u
niquely} positioned to be robust when diversity constraints are introduced, whil
e almost all other natural allocation rules fail this criterion. We also show th
at the guarantees achieved by Nash Welfare are nearly optimal within a widely st
udied class of allocation rules. We finally perform an empirical simulation on r
eal-world data that models ad allocations to show that this gap between Nash Wel
fare and other rules persists in the wild.
****
Activation Sharing with Asymmetric Paths Solves Weight Transport Problem without
Bidirectional Connection
Sunghyeon Woo, Jeongwoo Park, Jiwoo Hong, Dongsuk Jeon
One of the reasons why it is difficult for the brain to perform backpropagation
(BP) is the weight transport problem, which argues forward and feedback neurons
cannot share the same synaptic weights during learning in biological neural netw
orks. Recently proposed algorithms address the weight transport problem while pr
oviding good performance similar to BP in large-scale networks. However, they re
quire bidirectional connections between the forward and feedback neurons to trai
n their weights, which is observed to be rare in the biological brain. In this w
ork, we propose an Activation Sharing algorithm that removes the need for bidire
ctional connections between the two types of neurons. In this algorithm, hidden
layer outputs (activations) are shared across multiple layers during weight upda
tes. By applying this learning rule to both forward and feedback networks, we so
lve the weight transport problem without the constraint of bidirectional connect
ions, also achieving good performance even on deep convolutional neural networks
for various datasets. In addition, our algorithm could significantly reduce mem
ory access overhead when implemented in hardware.
****
BlendGAN: Implicitly GAN Blending for Arbitrary Stylized Face Generation
Mingcong Liu, Qiang Li, Zekui Qin, Guoxin Zhang, Pengfei Wan, Wen Zheng
Generative Adversarial Networks (GANs) have made a dramatic leap in high-fidelit
y image synthesis and stylized face generation. Recently, a layer-swapping mecha
nism has been developed to improve the stylization performance. However, this me
thod is incapable of fitting arbitrary styles in a single model and requires hun
dreds of style-consistent training images for each style. To address the above i
ssues, we propose BlendGAN for arbitrary stylized face generation by leveraging
a flexible blending strategy and a generic artistic dataset. Specifically, we fi
rst train a self-supervised style encoder on the generic artistic dataset to ext

ract the representations of arbitrary styles. In addition, a weighted blending module (WBM) is proposed to blend face and style representations implicitly and control the arbitrary stylization effect. By doing so, BlendGAN can gracefully fit arbitrary styles in a unified model while avoiding case-by-case preparation of style-consistent training images. To this end, we also present a novel large-scale artistic face dataset AAHQ. Extensive experiments demonstrate that BlendGAN outperforms state-of-the-art methods in terms of visual quality and style diversity for both latent-guided and reference-guided stylized face synthesis.
****
Differentially Private Model Personalization
Prateek Jain, John Rush, 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-authors prior to requesting a name change in the electronic proceedings.
****
Rates of Estimation of Optimal Transport Maps using Plug-in Estimators via Barycentric Projections
NABARUN DEB, Promit Ghosal, Bodhisattva Sen
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 Generalization despite Distribution Shift via Minimum Discriminating Information
Tobias Sutter, Andreas Krause, Daniel Kuhn
Training models that perform well under distribution shifts is a central challenge in machine learning. In this paper, we introduce a modeling framework where, in addition to training data, we have partial structural knowledge of the shifted test distribution. We employ the principle of minimum discriminating information to embed the available prior knowledge, and use distributionally robust optimization to account for uncertainty due to the limited samples. By leveraging large deviation results, we obtain explicit generalization bounds with respect to the unknown shifted distribution. Lastly, we demonstrate the versatility of our framework by demonstrating it on two rather distinct applications: (1) training classifiers on systematically biased data and (2) off-policy evaluation in Markov Decision Processes.
****
Soft Calibration Objectives for Neural Networks
Archit Karandikar, Nicholas Cain, Dustin Tran, Balaji Lakshminarayanan, Jonathon Shlens, Michael C. Mozer, Becca Roelofs
Optimal decision making requires that classifiers produce uncertainty estimates consistent with their empirical accuracy. However, deep neural networks are often under- or over-confident in their predictions. Consequently, methods have been developed to improve the calibration of their predictive uncertainty both during training and post-hoc. In this work, we propose differentiable losses to improve calibration based on a soft (continuous) version of the binning operation underlying popular calibration-error estimators. When incorporated into training, these soft calibration losses achieve state-of-the-art single-model ECE across multiple datasets with less than 1% decrease in accuracy. For instance, we observe an 82% reduction in ECE (70% relative to the post-hoc rescaled ECE) in exchange for a 0.7% relative decrease in accuracy relative to the cross entropy baseline on CIFAR-100.When incorporated post-training, the soft-binning-based calibration error objective improves upon temperature scaling, a popular recalibration method. Overall, experiments across losses and datasets demonstrate that using calibration-sensitive procedures yield better uncertainty estimates under dataset shift than the standard practice of using a cross entropy loss and post-hoc recalibration methods.
****

Distributional Gradient Matching for Learning Uncertain Neural Dynamics Models
Lenart Treven, Philippe Wenk, Florian Dorfler, Andreas Krause
Differential equations in general and neural ODEs in particular are an essential technique in continuous-time system identification. While many deterministic learning algorithms have been designed based on numerical integration via the adjoint method, many downstream tasks such as active learning, exploration in reinforcement learning, robust control, or filtering require accurate estimates of predictive uncertainties. In this work, we propose a novel approach towards estimating epistemically uncertain neural ODEs, avoiding the numerical integration bottleneck. Instead of modeling uncertainty in the ODE parameters, we directly model uncertainties in the state space. Our algorithm distributional gradient matching (DGM) jointly trains a smoother and a dynamics model and matches their gradients via minimizing a Wasserstein loss. Our experiments show that, compared to traditional approximate inference methods based on numerical integration, our approach is faster to train, faster at predicting previously unseen trajectories, and in the context of neural ODEs, significantly more accurate.
****
Shaping embodied agent behavior with activity-context priors from egocentric video
Tushar Nagarajan, Kristen Grauman
Complex physical tasks entail a sequence of object interactions, each with its own preconditions -- which can be difficult for robotic agents to learn efficiently solely through their own experience. We introduce an approach to discover activity-context priors from in-the-wild egocentric video captured with human worn cameras. For a given object, an activity-context prior represents the set of other compatible objects that are required for activities to succeed (e.g., a knife and cutting board brought together with a tomato are conducive to cutting). We encode our video-based prior as an auxiliary reward function that encourages an agent to bring compatible objects together before attempting an interaction. In this way, our model translates everyday human experience into embodied agent skills. We demonstrate our idea using egocentric EPIC-Kitchens video of people performing unscripted kitchen activities to benefit virtual household robotic agents performing various complex tasks in AI2-iTHOR, significantly accelerating agent learning.
****
Adjusting for Autocorrelated Errors in Neural Networks for Time Series
Fan-Keng Sun, Chris Lang, Duane Boning
An increasing body of research focuses on using neural networks to model time series. A common assumption in training neural networks via maximum likelihood estimation on time series is that the errors across time steps are uncorrelated. However, errors are actually autocorrelated in many cases due to the temporality of the data, which makes such maximum likelihood estimations inaccurate. In this paper, in order to adjust for autocorrelated errors, we propose to learn the autocorrelation coefficient jointly with the model parameters. In our experiments, we verify the effectiveness of our approach on time series forecasting. Results across a wide range of real-world datasets with various state-of-the-art models show that our method enhances performance in almost all cases. Based on these results, we suggest empirical critical values to determine the severity of autocorrelated errors. We also analyze several aspects of our method to demonstrate its advantages. Finally, other time series tasks are also considered to validate that our method is not restricted to only forecasting.
****
A Geometric Analysis of Neural Collapse with Unconstrained Features
Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, Qing Qu
We provide the first global optimization landscape analysis of Neural Collapse -- an intriguing empirical phenomenon that arises in the last-layer classifiers and features of neural networks during the terminal phase of training. As recently reported by Papyan et al., this phenomenon implies that (i) the class means and the last-layer classifiers all collapse to the vertices of a Simplex Equiangul

ar Tight Frame (ETF) up to scaling, and (ii) cross-example within-class variability of last-layer activations collapses to zero. We study the problem based on a simplified unconstrained feature model, which isolates the topmost layers from the classifier of the neural network. In this context, we show that the classical cross-entropy loss with weight decay has a benign global landscape, in the sense that the only global minimizers are the Simplex ETFs while all other critical points are strict saddles whose Hessian exhibit negative curvature directions. Our analysis of the simplified model not only explains what kind of features are learned in the last layer, but also shows why they can be efficiently optimized, matching the empirical observations in practical deep network architectures. These findings provide important practical implications. As an example, our experiments demonstrate that one may set the feature dimension equal to the number of classes and fix the last-layer classifier to be a Simplex ETF for network training, which reduces memory cost by over 20% on ResNet18 without sacrificing the generalization performance. The source code is available at https://github.com/tding1/Neural-Collapse.
****
NeRS: Neural Reflectance Surfaces for Sparse-view 3D Reconstruction in the Wild
Jason Zhang, Gengshan Yang, Shubham Tulsiani, Deva Ramanan
Recent history has seen a tremendous growth of work exploring implicit representations of geometry and radiance, popularized through Neural Radiance Fields (NeRF). Such works are fundamentally based on a (implicit) {\em volumetric} representation of occupancy, allowing them to model diverse scene structure including translucent objects and atmospheric obscurants. But because the vast majority of real-world scenes are composed of well-defined surfaces, we introduce a {\em surface} analog of such implicit models called Neural Reflectance Surfaces (NeRS). NeRS learns a neural shape representation of a closed surface that is diffeomorphic to a sphere, guaranteeing water-tight reconstructions. Even more importantly, surface parameterizations allow NeRS to learn (neural) bidirectional surface reflectance functions (BRDFs) that factorize view-dependent appearance into environmental illumination, diffuse color (albedo), and specular “shininess.” Finally, rather than illustrating our results on synthetic scenes or controlled in-the-lab capture, we assemble a novel dataset of multi-view images from online marketplaces for selling goods. Such “in-the-wild” multi-view image sets pose a number of challenges, including a small number of views with unknown/rough camera estimates. We demonstrate that surface-based neural reconstructions enable learning from such data, outperforming volumetric neural rendering-based reconstructions. We hope that NeRS serves as a first step toward building scalable, high-quality libraries of real-world shape, materials, and illumination.
****
Unleashing the Power of Contrastive Self-Supervised Visual Models via Contrast-Regularized Fine-Tuning
Yifan Zhang, Bryan Hooi, Dapeng Hu, Jian Liang, Jiashi Feng
Contrastive self-supervised learning (CSL) has attracted increasing attention for model pre-training via unlabeled data. The resulted CSL models provide instance-discriminative visual features that are uniformly scattered in the feature space. During deployment, the common practice is to directly fine-tune CSL models with cross-entropy, which however may not be the best strategy in practice. Although cross-entropy tends to separate inter-class features, the resulting models still have limited capability for reducing intra-class feature scattering that exists in CSL models. In this paper, we investigate whether applying contrastive learning to fine-tuning would bring further benefits, and analytically find that optimizing the contrastive loss benefits both discriminative representation learning and model optimization during fine-tuning. Inspired by these findings, we propose Contrast-regularized tuning (Core-tuning), a new approach for fine-tuning CSL models. Instead of simply adding the contrastive loss to the objective of fine-tuning, Core-tuning further applies a novel hard pair mining strategy for more effective contrastive fine-tuning, as well as smoothing the decision boundary to better exploit the learned discriminative feature space. Extensive experiments on image classification and semantic segmentation verify the effectiveness o

f Core-tuning.
****
Discovery of Options via Meta-Learned Subgoals
Vivek Veeriah, Tom Zahavy, Matteo Hessel, Zhongwen Xu, Junhyuk Oh, Iurii Kemaev, Hado P. van Hasselt, David Silver, Satinder Singh
Temporal abstractions in the form of options have been shown to help reinforcement learning (RL) agents learn faster. However, despite prior work on this topic, the problem of discovering options through interaction with an environment remains a challenge. In this paper, we introduce a novel meta-gradient approach for discovering useful options in multi-task RL environments. Our approach is based on a manager-worker decomposition of the RL agent, in which a manager maximises rewards from the environment by learning a task-dependent policy over both a set of task-independent discovered-options and primitive actions. The option-reward and termination functions that define a subgoal for each option are parameterised as neural networks and trained via meta-gradients to maximise their usefulness. Empirical analysis on gridworld and DeepMind Lab tasks show that: (1) our approach can discover meaningful and diverse temporally-extended options in multi-task RL domains, (2) the discovered options are frequently used by the agent while learning to solve the training tasks, and (3) that the discovered options help a randomly initialised manager learn faster in completely new tasks.
****
Near-Optimal Lower Bounds For Convex Optimization For All Orders of Smoothness
Ankit Garg, Robin Kothari, Praneeth Netrapalli, Suhail Sherif
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Topology-Imbalance Learning for Semi-Supervised Node Classification
Deli Chen, Yankai Lin, Guangxiang Zhao, Xuancheng Ren, Peng Li, Jie Zhou, Xu Sun
The class imbalance problem, as an important issue in learning node representations, has drawn increasing attention from the community. Although the imbalance considered by existing studies roots from the unequal quantity of labeled examples in different classes (quantity imbalance), we argue that graph data expose a unique source of imbalance from the asymmetric topological properties of the labeled nodes, i.e., labeled nodes are not equal in terms of their structural role in the graph (topology imbalance). In this work, we first probe the previously unknown topology-imbalance issue, including its characteristics, causes, and threats to semisupervised node classification learning. We then provide a unified view to jointly analyzing the quantity- and topology- imbalance issues by considering the node influence shift phenomenon with the Label Propagation algorithm. In light of our analysis, we devise an influence conflict detection–based metric Totoro to measure the degree of graph topology imbalance and propose a model-agnostic method ReNode to address the topology-imbalance issue by re-weighting the influence of labeled nodes adaptively based on their relative positions to class boundaries. Systematic experiments demonstrate the effectiveness and generalizability of our method in relieving topology-imbalance issue and promoting semi-supervised node classification. The further analysis unveils varied sensitivity of different graph neural networks (GNNs) to topology imbalance, which may serve as a new perspective in evaluating GNN architectures.
****
Gradient Inversion with Generative Image Prior
Jinwoo Jeon, jaechang Kim, Kangwook Lee, Sewoong Oh, Jungseul Ok
Federated Learning (FL) is a distributed learning framework, in which the local data never leaves clients’ devices to preserve privacy, and the server trains models on the data via accessing only the gradients of those local data. Without further privacy mechanisms such as differential privacy, this leaves the system vulnerable against an attacker who inverts those gradients to reveal clients’ sensitive data. However, a gradient is often insufficient to reconstruct the user data without any prior knowledge. By exploiting a generative model pretrained on

the data distribution, we demonstrate that data privacy can be easily breached.
Further, when such prior knowledge is unavailable, we investigate the possibilit
y of learning the prior from a sequence of gradients seen in the process of FL t
raining. We experimentally show that the prior in a form of generative model is
learnable from iterative interactions in FL. Our findings demonstrate that addit
ional mechanisms are necessary to prevent privacy leakage in FL.
****
Beta-CROWN: Efficient Bound Propagation with Per-neuron Split Constraints for Ne
ural Network Robustness Verification
Shiqi Wang, Huan Zhang, Kaidi Xu, Xue Lin, Suman Jana, Cho-Jui Hsieh, J. Zico Ko
lter
Requests for name changes in the electronic proceedings will be accepted with 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.
****
Autobahn: Automorphism-based Graph Neural Nets
Erik Thiede, Wenda Zhou, Risi Kondor
We introduce Automorphism-based graph neural networks (Autobahn), a new family o
f graph neural networks. In an Autobahn, we decompose the graph into a collectio
n of subgraphs and apply local convolutions that are equivariant to each subgrap
h's automorphism group. Specific choices of local neighborhoods and subgraphs re
cover existing architectures such as message passing neural networks. Our formal
ism also encompasses novel architectures: as an example, we introduce a graph ne
ural network that decomposes the graph into paths and cycles. The resulting conv
olutions reflect the natural way that parts of the graph can transform, preservi
ng the intuitive meaning of convolution without sacrificing global permutation e
quivariance. We validate our approach by applying Autobahn to molecular graphs,
where it achieves results competitive with state-of-the-art message passing algo
rithms.
****
Data Augmentation Can Improve Robustness
Sylvestre-Alvise Rebuffi, Sven Gowal, Dan Andrei Calian, Florian Stimberg, Olivi
a Wiles, Timothy A Mann
Requests for name changes in the electronic proceedings will be accepted with 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 Explicit Duration Switching Models for Time Series
Abdul Fatir Ansari, Konstantinos Benidis, Richard Kurle, Ali Caner Turkmen, Haro
ld Soh, Alexander J. Smola, Bernie Wang, Tim Januschowski
Many complex time series can be effectively subdivided into distinct regimes tha
t exhibit persistent dynamics. Discovering the switching behavior and the statis
tical patterns in these regimes is important for understanding the underlying dy
namical system. We propose the Recurrent Explicit Duration Switching Dynamical S
ystem (RED-SDS), a flexible model that is capable of identifying both state- and
 time-dependent switching dynamics. State-dependent switching is enabled by a re
current state-to-switch connection and an explicit duration count variable is us
ed to improve the time-dependent switching behavior. We demonstrate how to perfo
rm efficient inference using a hybrid algorithm that approximates the posterior
of the continuous states via an inference network and performs exact inference f
or the discrete switches and counts. The model is trained by maximizing a Monte
Carlo lower bound of the marginal log-likelihood that can be computed efficientl
y as a byproduct of the inference routine. Empirical results on multiple dataset
s demonstrate that RED-SDS achieves considerable improvement in time series seg
mentation and competitive forecasting performance against the state of the art.
****
Shared Independent Component Analysis for Multi-Subject Neuroimaging
Hugo Richard, Pierre Ablin, Bertrand Thirion, Alexandre Gramfort, Aapo Hyvarinen

We consider shared response modeling, a multi-view learning problem where one wants to identify common components from multiple datasets or views. We introduce Shared Independent Component Analysis (ShICA) that models eachview as a linear transform of shared independent components contaminated by additive Gaussian noise. We show that this model is identifiable if the components are either non-Gaussian or have enough diversity in noise variances. We then show that in some cases multi-set canonical correlation analysis can recover the correct unmixing matrices, but that even a small amount of sampling noise makes Multiset CCA fail. To solve this problem, we propose to use joint diagonalization after Multiset CCA, leading to a new approach called ShICA-J. We show via simulations that ShICA-J leads to improved results while being very fast to fit. While ShICA-J is based on second-order statistics, we further propose to leverage non-Gaussianity of the components using a maximum-likelihood method, ShICA-ML, that is both more accurate and more costly. Further, ShICA comes with a principled method for shared components estimation. Finally, we provide empirical evidence on fMRI and MEG datasets that ShICA yields more accurate estimation of the componentsthan alternatives.
****
Shape from Blur: Recovering Textured 3D Shape and Motion of Fast Moving Objects
Denys Rozumnyi, Martin R. Oswald, Vittorio Ferrari, Marc Pollefeys
We address the novel task of jointly reconstructing the 3D shape, texture, and motion of an object from a single motion-blurred image. While previous approaches address the deblurring problem only in the 2D image domain, our proposed rigorous modeling of all object properties in the 3D domain enables the correct description of arbitrary object motion. This leads to significantly better image decomposition and sharper deblurring results. We model the observed appearance of a motion-blurred object as a combination of the background and a 3D object with constant translation and rotation. Our method minimizes a loss on reconstructing the input image via differentiable rendering with suitable regularizers. This enables estimating the textured 3D mesh of the blurred object with high fidelity. Our method substantially outperforms competing approaches on several benchmarks for fast moving objects deblurring. Qualitative results show that the reconstructed 3D mesh generates high-quality temporal super-resolution and novel views of the deblurred object.
****
Batched Thompson Sampling
Cem Kalkanli, Ayfer Ozgur
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Delayed Gradient Averaging: Tolerate the Communication Latency for Federated Learning
Ligeng Zhu, Hongzhou Lin, Yao Lu, Yujun Lin, Song Han
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Focal Attention for Long-Range Interactions in Vision Transformers
Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, Jianfeng Gao
Recently, Vision Transformer and its variants have shown great promise on various computer vision tasks. The ability to capture local and global visual dependencies through self-attention is the key to its success. But it also brings challenges due to quadratic computational overhead, especially for the high-resolution vision tasks(e.g., object detection). Many recent works have attempted to reduce the cost and improve model performance by applying either coarse-grained global attention or fine-grained local attention. However, both approaches cripple th

e modeling power of the original self-attention mechanism of multi-layer Transformers, leading to sub-optimal solutions. In this paper, we present focal attention, a new attention mechanism that incorporates both fine-grained local and coarse-grained global interactions. In this new mechanism, each token attends its closest surrounding tokens at the fine granularity and the tokens far away at a coarse granularity and thus can capture both short- and long-range visual dependencies efficiently and effectively. With focal attention, we propose a new variant of Vision Transformer models, called Focal Transformers, which achieve superior performance over the state-of-the-art (SoTA) Vision Transformers on a range of public image classification and object detection benchmarks. In particular, our Focal Transformer models with a moderate size of 51.1M and a large size of 89.8M achieve 83.6% and 84.0%Top-1 accuracy, respectively, on ImageNet classification at 224×224. When employed as the backbones, Focal Transformers achieve consistent and substantial improvements over the current SoTA Swin Transformers [44] across 6 different object detection methods. Our largest Focal Transformer yields58.7/59.0boxmAPs and50.9/51.3mask mAPs on COCO mini-val/test-dev, and55.4mIoU onADE20K for semantic segmentation, creating new SoTA on three of the most challenging computer vision tasks.
****

Scalable and Stable Surrogates for Flexible Classifiers with Fairness Constraints

Henry C Bendekgey, Erik Sudderth

We investigate how fairness relaxations scale to flexible classifiers like deep neural networks for images and text. We analyze an easy-to-use and robust way of imposing fairness constraints when training, and through this framework prove that some prior fairness surrogates exhibit degeneracies for non-convex models. We resolve these problems via three new surrogates: an adaptive data re-weighting, and two smooth upper-bounds that are provably more robust than some previous methods. Our surrogates perform comparably to the state-of-the-art on low-dimensional fairness benchmarks, while achieving superior accuracy and stability for more complex computer vision and natural language processing tasks.
****

Residual Pathway Priors for Soft Equivariance Constraints

Marc Finzi, Gregory Benton, Andrew G. Wilson

Models such as convolutional neural networks restrict the hypothesis space to a set of functions satisfying equivariance constraints, and improve generalization in problems by capturing relevant symmetries. However, symmetries are often only partially respected, preventing models with restriction biases from fitting the data. We introduce Residual Pathway Priors (RPPs) as a method for converting hard architectural constraints into soft priors, guiding models towards structured solutions while retaining the ability to capture additional complexity. RPPs are resilient to approximate or misspecified symmetries, and are as effective as fully constrained models even when symmetries are exact. We show that RPPs provide compelling performance on both model-free and model-based reinforcement learning problems, where contact forces and directional rewards violate the assumptions of equivariant networks. Finally, we demonstrate that RPPs have broad applicability, including dynamical systems, regression, and classification.
****

Optimal Algorithms for Stochastic Contextual Preference Bandits

Aadirupa Saha

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

Tight High Probability Bounds for Linear Stochastic Approximation with Fixed Stepsize

Alain Durmus, Eric Moulines, Alexey Naumov, Sergey Samsonov, Kevin Scaman, 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-authors prior to requesting a name change in the electronic proceedings.
****

Learning Large Neighborhood Search Policy for Integer Programming
Yaoxin Wu, Wen Song, Zhiguang Cao, Jie Zhang
We propose a deep reinforcement learning (RL) method to learn large neighborhood search (LNS) policy for integer programming (IP). The RL policy is trained as the destroy operator to select a subset of variables at each step, which is reoptimized by an IP solver as the repair operator. However, the combinatorial number of variable subsets prevents direct application of typical RL algorithms. To tackle this challenge, we represent all subsets by factorizing them into binary decisions on each variable. We then design a neural network to learn policies for each variable in parallel, trained by a customized actor-critic algorithm. We evaluate the proposed method on four representative IP problems. Results show that it can find better solutions than SCIP in much less time, and significantly outperform other LNS baselines with the same runtime. Moreover, these advantages notably persist when the policies generalize to larger problems. Further experiments with Gurobi also reveal that our method can outperform this state-of-the-art commercial solver within the same time limit.
****

Dynamic Trace Estimation
Prathamesh Dharangutte, Christopher Musco
We study a dynamic version of the implicit trace estimation problem. Given access to an oracle for computing matrix-vector multiplications with a dynamically changing matrix A, our goal is to maintain an accurate approximation to A's trace using as few multiplications as possible. We present a practical algorithm for solving this problem and prove that, in a natural setting, its complexity is quadratically better than the standard solution of repeatedly applying Hutchinson's stochastic trace estimator. We also provide an improved algorithm assuming additional common assumptions on A's dynamic updates. We support our theory with empirical results, showing significant computational improvements on three applications in machine learning and network science: tracking moments of the Hessian spectral density during neural network optimization, counting triangles and estimating natural connectivity in a dynamically changing graph.
****

Provable Representation Learning for Imitation with Contrastive Fourier Features
Ofir Nachum, Mengjiao Yang
In imitation learning, it is common to learn a behavior policy to match an unknown target policy via max-likelihood training on a collected set of target demonstrations. In this work, we consider using offline experience datasets -- potentially far from the target distribution -- to learn low-dimensional state representations that provably accelerate the sample-efficiency of downstream imitation learning. A central challenge in this setting is that the unknown target policy itself may not exhibit low-dimensional behavior, and so there is a potential for the representation learning objective to alias states in which the target policy acts differently. Circumventing this challenge, we derive a representation learning objective that provides an upper bound on the performance difference between the target policy and a low-dimensional policy trained with max-likelihood, and this bound is tight regardless of whether the target policy itself exhibits low-dimensional structure. Moving to the practicality of our method, we show that our objective can be implemented as contrastive learning, in which the transition dynamics are approximated by either an implicit energy-based model or, in some special cases, an implicit linear model with representations given by random Fourier features. Experiments on both tabular environments and high-dimensional Atari games provide quantitative evidence for the practical benefits of our proposed objective.
****

MICo: Improved representations via sampling-based state similarity for Markov decision processes

Pablo Samuel Castro, Tyler Kastner, Prakash Panangaden, Mark Rowland
We present a new behavioural distance over the state space of a Markov decision process, and demonstrate the use of this distance as an effective means of shaping the learnt representations of deep reinforcement learning agents. While existing notions of state similarity are typically difficult to learn at scale due to high computational cost and lack of sample-based algorithms, our newly-proposed distance addresses both of these issues. In addition to providing detailed theoretical analyses, we provide empirical evidence that learning this distance alongside the value function yields structured and informative representations, including strong results on the Arcade Learning Environment benchmark.
****
Counterfactual Explanations in Sequential Decision Making Under Uncertainty
Stratis Tsirtsis, Abir De, Manuel Rodriguez
Methods to find counterfactual explanations have predominantly focused on one-step decision making processes. In this work, we initiate the development of methods to find counterfactual explanations for decision making processes in which multiple, dependent actions are taken sequentially over time. We start by formally characterizing a sequence of actions and states using finite horizon Markov decision processes and the Gumbel-Max structural causal model. Building upon this characterization, we formally state the problem of finding counterfactual explanations for sequential decision making processes. In our problem formulation, the counterfactual explanation specifies an alternative sequence of actions differing in at most k actions from the observed sequence that could have led the observed process realization to a better outcome. Then, we introduce a polynomial time algorithm based on dynamic programming to build a counterfactual policy that is guaranteed to always provide the optimal counterfactual explanation on every possible realization of the counterfactual environment dynamics. We validate our algorithm using both synthetic and real data from cognitive behavioral therapy and show that the counterfactual explanations our algorithm finds can provide valuable insights to enhance sequential decision making under uncertainty.
****
Streaming Linear System Identification with Reverse Experience Replay
Suhas Kowshik, Dheeraj Nagaraj, Prateek Jain, Praneeth Netrapalli
We consider the problem of estimating a linear time-invariant (LTI) dynamical system from a single trajectory via streaming algorithms, which is encountered in several applications including reinforcement learning (RL) and time-series analysis. While the LTI system estimation problem is well-studied in the {\em offline} setting, the practically important streaming/online setting has received little attention. Standard streaming methods like stochastic gradient descent (SGD) are unlikely to work since streaming points can be highly correlated. In this work, we propose a novel streaming algorithm, SGD with Reverse Experience Replay (SGD-RER), that is inspired by the experience replay (ER) technique popular in the RL literature. SGD-RER divides data into small buffers and runs SGD backwards on the data stored in the individual buffers. We show that this algorithm exactly deconstructs the dependency structure and obtains information theoretically optimal guarantees for both parameter error and prediction error. Thus, we provide the first -- to the best of our knowledge -- optimal SGD-style algorithm for the classical problem of linear system identification with a first order oracle. Furthermore, SGD-RER can be applied to more general settings like sparse LTI identification with known sparsity pattern, and non-linear dynamical systems. Our work demonstrates that the knowledge of data dependency structure can aid us in designing statistically and computationally efficient algorithms which can ``decorrelate'' streaming samples.
****
SmoothMix: Training Confidence-calibrated Smoothed Classifiers for Certified Robustness
Jongheon Jeong, Sejun Park, Minkyu Kim, Heung-Chang Lee, Do-Guk Kim, Jinwoo Shin
Requests for name changes in the electronic proceedings will be accepted with 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.
****
Action-guided 3D Human Motion Prediction
Jiangxin Sun, Zihang Lin, Xintong Han, Jian-Fang Hu, Jia Xu, Wei-Shi Zheng
The ability of forecasting future human motion is important for human-machine interaction systems to understand human behaviors and make interaction. In this work, we focus on developing models to predict future human motion from past observed video frames. Motivated by the observation that human motion is closely related to the action being performed, we propose to explore action context to guide motion prediction. Specifically, we construct an action-specific memory bank to store representative motion dynamics for each action category, and design a query-read process to retrieve some motion dynamics from the memory bank. The retrieved dynamics are consistent with the action depicted in the observed video frames and serve as a strong prior knowledge to guide motion prediction. We further formulate an action constraint loss to ensure the global semantic consistency of the predicted motion. Extensive experiments demonstrate the effectiveness of the proposed approach, and we achieve state-of-the-art performance on 3D human motion prediction.
****
Meta-Learning the Search Distribution of Black-Box Random Search Based Adversarial Attacks
Maksym Yatsura, Jan Metzen, Matthias Hein
Adversarial attacks based on randomized search schemes have obtained state-of-the-art results in black-box robustness evaluation recently. However, as we demonstrate in this work, their efficiency in different query budget regimes depends on manual design and heuristic tuning of the underlying proposal distributions. We study how this issue can be addressed by adapting the proposal distribution online based on the information obtained during the attack. We consider Square Attack, which is a state-of-the-art score-based black-box attack, and demonstrate how its performance can be improved by a learned controller that adjusts the parameters of the proposal distribution online during the attack. We train the controller using gradient-based end-to-end training on a CIFAR10 model with white box access. We demonstrate that plugging the learned controller into the attack consistently improves its black-box robustness estimate in different query regimes by up to 20% for a wide range of different models with black-box access. We further show that the learned adaptation principle transfers well to the other data distributions such as CIFAR100 or ImageNet and to the targeted attack setting.
****
Validating the Lottery Ticket Hypothesis with Inertial Manifold Theory
Zeru Zhang, Jiayin Jin, Zijie Zhang, Yang Zhou, Xin Zhao, Jiaxiang Ren, Ji Liu, Lingfei Wu, Ruoming Jin, Dejing Dou
Despite achieving remarkable efficiency, traditional network pruning techniques often follow manually-crafted heuristics to generate pruned sparse networks. Such heuristic pruning strategies are hard to guarantee that the pruned networks achieve test accuracy comparable to the original dense ones. Recent works have empirically identified and verified the Lottery Ticket Hypothesis (LTH): a randomly-initialized dense neural network contains an extremely sparse subnetwork, which can be trained to achieve similar accuracy to the former. Due to the lack of theoretical evidence, they often need to run multiple rounds of expensive training and pruning over the original large networks to discover the sparse subnetworks with low accuracy loss. By leveraging dynamical systems theory and inertial manifold theory, this work theoretically verifies the validity of the LTH. We explore the possibility of theoretically lossless pruning as well as one-time pruning, compared with existing neural network pruning and LTH techniques. We reformulate the neural network optimization problem as a gradient dynamical system and reduce this high-dimensional system onto inertial manifolds to obtain a low-dimensional system regarding pruned subnetworks. We demonstrate the precondition and existence of pruned subnetworks and prune the original networks in terms of the gap in their spectrum that make the subnetworks have the smallest dimensions.
****

Are My Deep Learning Systems Fair? An Empirical Study of Fixed-Seed Training
Shangshu Qian, Viet Hung Pham, Thibaud Lutellier, Zeou Hu, Jungwon Kim, Lin Tan, Yaoliang Yu, Jiahao Chen, Sameena Shah
Deep learning (DL) systems have been gaining popularity in critical tasks such as credit evaluation and crime prediction. Such systems demand fairness. Recent work shows that DL software implementations introduce variance: identical DL training runs (i.e., identical network, data, configuration, software, and hardware) with a fixed seed produce different models. Such variance could make DL models and networks violate fairness compliance laws, resulting in negative social impact. In this paper, we conduct the first empirical study to quantify the impact of software implementation on the fairness and its variance of DL systems. Our study of 22 mitigation techniques and five baselines reveals up to 12.6% fairness variance across identical training runs with identical seeds. In addition, most debiasing algorithms have a negative impact on the model such as reducing model accuracy, increasing fairness variance, or increasing accuracy variance. Our literature survey shows that while fairness is gaining popularity in artificial intelligence (AI) related conferences, only 34.4% of the papers use multiple identical training runs to evaluate their approach, raising concerns about their results' validity. We call for better fairness evaluation and testing protocols to improve fairness and fairness variance of DL systems as well as DL research validity and reproducibility at large.
****
Rectangular Flows for Manifold Learning
Anthony L. Caterini, Gabriel Loaiza-Ganem, Geoff Pleiss, John P. Cunningham
Normalizing flows are invertible neural networks with tractable change-of-volume terms, which allow optimization of their parameters to be efficiently performed via maximum likelihood. However, data of interest are typically assumed to live in some (often unknown) low-dimensional manifold embedded in a high-dimensional ambient space. The result is a modelling mismatch since -- by construction -- the invertibility requirement implies high-dimensional support of the learned distribution. Injective flows, mappings from low- to high-dimensional spaces, aim to fix this discrepancy by learning distributions on manifolds, but the resulting volume-change term becomes more challenging to evaluate. Current approaches either avoid computing this term entirely using various heuristics, or assume the manifold is known beforehand and therefore are not widely applicable. Instead, we propose two methods to tractably calculate the gradient of this term with respect to the parameters of the model, relying on careful use of automatic differentiation and techniques from numerical linear algebra. Both approaches perform end-to-end nonlinear manifold learning and density estimation for data projected onto this manifold. We study the trade-offs between our proposed methods, empirically verify that we outperform approaches ignoring the volume-change term by more accurately learning manifolds and the corresponding distributions on them, and show promising results on out-of-distribution detection. Our code is available at https://github.com/layer6ai-labs/rectangular-flows.
****
On the Generative Utility of Cyclic Conditionals
Chang Liu, Haoyue Tang, Tao Qin, Jintao Wang, Tie-Yan Liu
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Structural Credit Assignment in Neural Networks using Reinforcement Learning
Dhawal Gupta, Gabor Mihucz, Matthew Schlegel, James Kostas, Philip S. Thomas, Martha White
Structural credit assignment in neural networks is a long-standing problem, with a variety of alternatives to backpropagation proposed to allow for local training of nodes. One of the early strategies was to treat each node as an agent and use a reinforcement learning method called REINFORCE to update each node locally with only a global reward signal. In this work, we revisit this approach and in

vestigate if we can leverage other reinforcement learning approaches to improve learning. We first formalize training a neural network as a finite-horizon reinforcement learning problem and discuss how this facilitates using ideas from reinforcement learning like off-policy learning. We show that the standard on-policy REINFORCE approach, even with a variety of variance reduction approaches, learns suboptimal solutions. We introduce an off-policy approach, to facilitate reasoning about the greedy action for other agents and help overcome stochasticity in other agents. We conclude by showing that these networks of agents can be more robust to correlated samples when learning online.
****
A Near-Optimal Algorithm for Stochastic Bilevel Optimization via Double-Momentum
Prashant Khanduri, Siliang Zeng, Mingyi Hong, Hoi-To Wai, Zhaoran Wang, Zhuoran 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.
****
Generalized Jensen-Shannon Divergence Loss for Learning with Noisy Labels
Erik Englesson, Hossein Azizpour
Prior works have found it beneficial to combine provably noise-robust loss functions e.g., mean absolute error (MAE) with standard categorical loss function e.g. cross entropy (CE) to improve their learnability. Here, we propose to use Jensen-Shannon divergence as a noise-robust loss function and show that it interestingly interpolate between CE and MAE with a controllable mixing parameter. Furthermore, we make a crucial observation that CE exhibit lower consistency around noisy data points. Based on this observation, we adopt a generalized version of the Jensen-Shannon divergence for multiple distributions to encourage consistency around data points. Using this loss function, we show state-of-the-art results on both synthetic (CIFAR), and real-world (e.g., WebVision) noise with varying noise rates.
****
Continual Learning via Local Module Composition
Oleksiy Ostapenko, Pau Rodriguez, Massimo Caccia, Laurent Charlin
Modularity is a compelling solution to continual learning (CL), the problem of modeling sequences of related tasks. Learning and then composing modules to solve different tasks provides an abstraction to address the principal challenges of CL including catastrophic forgetting, backward and forward transfer across tasks, and sub-linear model growth. We introduce local module composition (LMC), an approach to modular CL where each module is provided a local structural component that estimates a module’s relevance to the input. Dynamic module composition is performed layer-wise based on local relevance scores. We demonstrate that agnosticity to task identities (IDs) arises from (local) structural learning that is module-specific as opposed to the task- and/or model-specific as in previous works, making LMC applicable to more CL settings compared to previous works. In addition, LMC also tracks statistics about the input distribution and adds new modules when outlier samples are detected. In the first set of experiments, LMC performs favorably compared to existing methods on the recent Continual Transfer-learning Benchmark without requiring task identities. In another study, we show that the locality of structural learning allows LMC to interpolate to related but unseen tasks (OOD), as well as to compose modular networks trained independently on different task sequences into a third modular network without any fine-tuning. Finally, in search for limitations of LMC we study it on more challenging sequences of 30 and 100 tasks, demonstrating that local module selection becomes much more challenging in presence of a large number of candidate modules. In this setting best performing LMC spawns much fewer modules compared to an oracle based baseline, however, it reaches a lower overall accuracy. The codebase is available under https://github.com/oleksost/LMC.
****
Model-Based Episodic Memory Induces Dynamic Hybrid Controls

Hung Le, Thommen Karimpanal George, Majid Abdolshah, Truyen Tran, Svetha Venkatesh
Episodic control enables sample efficiency in reinforcement learning by recalling past experiences from an episodic memory. We propose a new model-based episodic memory of trajectories addressing current limitations of episodic control. Our memory estimates trajectory values, guiding the agent towards good policies. Built upon the memory, we construct a complementary learning model via a dynamic hybrid control unifying model-based, episodic and habitual learning into a single architecture. Experiments demonstrate that our model allows significantly faster and better learning than other strong reinforcement learning agents across a variety of environments including stochastic and non-Markovian settings.
****
FedDR – Randomized Douglas-Rachford Splitting Algorithms for Nonconvex Federated Composite Optimization
Quoc Tran Dinh, Nhan H Pham, Dzung Phan, Lam Nguyen
We develop two new algorithms, called, FedDR and asyncFedDR, for solving a fundamental nonconvex composite optimization problem in federated learning. Our algorithms rely on a novel combination between a nonconvex Douglas-Rachford splitting method, randomized block-coordinate strategies, and asynchronous im- plementation. They can also handle convex regularizers. Unlike recent methods in the literature, e.g., FedSplit and FedPD, our algorithms update only a subset of users at each communication round, and possibly in an asynchronous manner, making them more practical. These new algorithms can handle statistical and sys- tem heterogeneity, which are the two main challenges in federated learning, while achieving the best known communication complexity. In fact, our new algorithms match the communication complexity lower bound up to a constant factor under standard assumptions. Our numerical experiments illustrate the advantages of our methods over existing algorithms on synthetic and real datasets.
****
Adversarial Examples Make Strong Poisons
Liam Fowl, Micah Goldblum, Ping-yeh Chiang, Jonas Geiping, Wojciech Czaja, Tom Goldstein
The adversarial machine learning literature is largely partitioned into evasion attacks on testing data and poisoning attacks on training data. In this work, we show that adversarial examples, originally intended for attacking pre-trained models, are even more effective for data poisoning than recent methods designed specifically for poisoning. In fact, adversarial examples with labels re-assigned by the crafting network remain effective for training, suggesting that adversarial examples contain useful semantic content, just with the "wrong" labels (according to a network, but not a human). Our method, adversarial poisoning, is substantially more effective than existing poisoning methods for secure dataset release, and we release a poisoned version of ImageNet, ImageNet-P, to encourage research into the strength of this form of data obfuscation.
****
Coresets for Decision Trees of Signals
Ibrahim Jubran, Ernesto Evgeniy Sanches Shayda, Ilan I Newman, Dan Feldman
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Local plasticity rules can learn deep representations using self-supervised contrastive predictions
Bernd Illing, Jean Ventura, Guillaume Bellec, Wulfram Gerstner
Learning in the brain is poorly understood and learning rules that respect biological constraints, yet yield deep hierarchical representations, are still unknown. Here, we propose a learning rule that takes inspiration from neuroscience and recent advances in self-supervised deep learning. Learning minimizes a simple layer-specific loss function and does not need to back-propagate error signals within or between layers. Instead, weight updates follow a local, Hebbian, learnin

g rule that only depends on pre- and post-synaptic neuronal activity, predictive dendritic input and widely broadcasted modulation factors which are identical for large groups of neurons. The learning rule applies contrastive predictive learning to a causal, biological setting using saccades (i.e. rapid shifts in gaze direction). We find that networks trained with this self-supervised and local rule build deep hierarchical representations of images, speech and video.
****
MobTCast: Leveraging Auxiliary Trajectory Forecasting for Human Mobility Prediction
Hao Xue, Flora Salim, Yongli Ren, Nuria Oliver
Human mobility prediction is a core functionality in many location-based services and applications. However, due to the sparsity of mobility data, it is not an easy task to predict future POIs (place-of-interests) that are going to be visited. In this paper, we propose MobTCast, a Transformer-based context-aware network for mobility prediction. Specifically, we explore the influence of four types of context in mobility prediction: temporal, semantic, social, and geographical contexts. We first design a base mobility feature extractor using the Transformer architecture, which takes both the history POI sequence and the semantic information as input. It handles both the temporal and semantic contexts. Based on the base extractor and the social connections of a user, we employ a self-attention module to model the influence of the social context. Furthermore, unlike existing methods, we introduce a location prediction branch in MobTCast as an auxiliary task to model the geographical context and predict the next location. Intuitively, the geographical distance between the location of the predicted POI and the predicted location from the auxiliary branch should be as close as possible. To reflect this relation, we design a consistency loss to further improve the POI prediction performance. In our experimental results, MobTCast outperforms other state-of-the-art next POI prediction methods. Our approach illustrates the value of including different types of context in next POI prediction.
****
Early Convolutions Help Transformers See Better
Tete Xiao, Mannat Singh, Eric Mintun, Trevor Darrell, Piotr Dollar, Ross Girshick
Vision transformer (ViT) models exhibit substandard optimizability. In particular, they are sensitive to the choice of optimizer (AdamW vs. SGD), optimizer hyperparameters, and training schedule length. In comparison, modern convolutional neural networks are easier to optimize. Why is this the case? In this work, we conjecture that the issue lies with the patchify stem of ViT models, which is implemented by a stride-p p×p convolution (p = 16 by default) applied to the input image. This large-kernel plus large-stride convolution runs counter to typical design choices of convolutional layers in neural networks. To test whether this atypical design choice causes an issue, we analyze the optimization behavior of ViT models with their original patchify stem versus a simple counterpart where we replace the ViT stem by a small number of stacked stride-two 3×3 convolutions. While the vast majority of computation in the two ViT designs is identical, we find that this small change in early visual processing results in markedly different training behavior in terms of the sensitivity to optimization settings as well as the final model accuracy. Using a convolutional stem in ViT dramatically increases optimization stability and also improves peak performance (by ~1-2% top-1 accuracy on ImageNet-1k), while maintaining flops and runtime. The improvement can be observed across the wide spectrum of model complexities (from 1G to 36G flops) and dataset scales (from ImageNet-1k to ImageNet-21k). These findings lead us to recommend using a standard, lightweight convolutional stem for ViT models in this regime as a more robust architectural choice compared to the original ViT model design.
****
Error Compensated Distributed SGD Can Be Accelerated
Xun Qian, Peter Richtarik, Tong Zhang
Gradient compression is a recent and increasingly popular technique for reducing the communication cost in distributed training of large-scale machine learning

models. In this work we focus on developing efficient distributed methods that c
an work for any compressor satisfying a certain contraction property, which incl
udes both unbiased (after appropriate scaling) and biased compressors such as Ra
ndK and TopK. Applied naively, gradient compression introduces errors that eithe
r slow down convergence or lead to divergence. A popular technique designed to t
ackle this issue is error compensation/error feedback. Due to the difficulties a
ssociated with analyzing biased compressors, it is not known whether gradient co
mpression with error compensation can be combined with acceleration. In this wor
k, we show for the first time that error compensated gradient compression method
s can be accelerated. In particular, we propose and study the error compensated
loopless Katyusha method, and establish an accelerated linear convergence rate u
nder standard assumptions. We show through numerical experiments that the propos
ed method converges with substantially fewer communication rounds than previous
error compensated algorithms.
****
InfoGCL: Information-Aware Graph Contrastive Learning
Dongkuan Xu, Wei Cheng, Dongsheng Luo, Haifeng Chen, Xiang Zhang
Various graph contrastive learning models have been proposed to improve the perf
ormance of tasks on graph datasets in recent years. While effective and prevalen
t, these models are usually carefully customized. In particular, despite all rec
ent work create two contrastive views, they differ in a variety of view augmenta
tions, architectures, and objectives. It remains an open question how to build y
our graph contrastive learning model from scratch for particular graph tasks and
 datasets. In this work, we aim to fill this gap by studying how graph informati
on is transformed and transferred during the contrastive learning process, and p
roposing an information-aware graph contrastive learning framework called InfoGC
L. The key to the success of the proposed framework is to follow the Information
 Bottleneck principle to reduce the mutual information between contrastive parts
 while keeping task-relevant information intact at both the levels of the indivi
dual module and the entire framework so that the information loss during graph r
epresentation learning can be minimized. We show for the first time that all rec
ent graph contrastive learning methods can be unified by our framework. Based on
 theoretical and empirical analysis on benchmark graph datasets, we show that In
foGCL achieves state-of-the-art performance in the settings of both graph classi
fication and node classification tasks.
****
Meta-Learning for Relative Density-Ratio Estimation
Atsutoshi Kumagai, Tomoharu Iwata, Yasuhiro Fujiwara
The ratio of two probability densities, called a density-ratio, is a vital quant
ity in machine learning. In particular, a relative density-ratio, which is a bou
nded extension of the density-ratio, has received much attention due to its stab
ility and has been used in various applications such as outlier detection and da
taset comparison. Existing methods for (relative) density-ratio estimation (DRE)
 require many instances from both densities. However, sufficient instances are o
ften unavailable in practice. In this paper, we propose a meta-learning method f
or relative DRE, which estimates the relative density-ratio from a few instances
 by using knowledge in related datasets. Specifically, given two datasets that c
onsist of a few instances, our model extracts the datasets' information by using
 neural networks and uses it to obtain instance embeddings appropriate for the r
elative DRE. We model the relative density-ratio by a linear model on the embedd
ed space, whose global optimum solution can be obtained as a closed-form solutio
n. The closed-form solution enables fast and effective adaptation to a few insta
nces, and its differentiability enables us to train our model such that the expe
cted test error for relative DRE can be explicitly minimized after adapting to a
 few instances. We empirically demonstrate the effectiveness of the proposed met
hod by using three problems: relative DRE, dataset comparison, and outlier detec
tion.
****
Overcoming the curse of dimensionality with Laplacian regularization in semi-sup
ervised learning

Vivien Cabannes, Loucas Pillaud-Vivien, Francis Bach, Alessandro Rudi
As annotations of data can be scarce in large-scale practical problems, leveraging unlabelled examples is one of the most important aspects of machine learning. This is the aim of semi-supervised learning. To benefit from the access to unlabelled data, it is natural to diffuse smoothly knowledge of labelled data to unlabelled one. This induces to the use of Laplacian regularization. Yet, current implementations of Laplacian regularization suffer from several drawbacks, notably the well-known curse of dimensionality. In this paper, we design a new class of algorithms overcoming this issue, unveiling a large body of spectral filtering methods. Additionally, we provide a statistical analysis showing that our estimators exhibit desirable behaviors. They are implemented through (reproducing) kernel methods, for which we provide realistic computational guidelines in order to make our method usable with large amounts of data.
****
Unlabeled Principal Component Analysis
Yunzhen Yao, Liangzu Peng, Manolis Tsakiris
We introduce robust principal component analysis from a data matrix in which the entries of its columns have been corrupted by permutations, termed Unlabeled Principal Component Analysis (UPCA). Using algebraic geometry, we establish that UPCA is a well-defined algebraic problem in the sense that the only matrices of minimal rank that agree with the given data are row-permutations of the ground-truth matrix, arising as the unique solutions of a polynomial system of equations. Further, we propose an efficient two-stage algorithmic pipeline for UPCA suitable for the practically relevant case where only a fraction of the data have been permuted. Stage-I employs outlier-robust PCA methods to estimate the ground-truth column-space. Equipped with the column-space, Stage-II applies recent methods for unlabeled sensing to restore the permuted data. Experiments on synthetic data, face images, educational and medical records reveal the potential of UPCA for applications such as data privatization and record linkage.
****
Causal-BALD: Deep Bayesian Active Learning of Outcomes to Infer Treatment-Effects from Observational Data
Andrew Jesson, Panagiotis Tigas, Joost van Amersfoort, Andreas Kirsch, Uri Shalit, Yarin Gal
Estimating personalized treatment effects from high-dimensional observational data is essential in situations where experimental designs are infeasible, unethical, or expensive. Existing approaches rely on fitting deep models on outcomes observed for treated and control populations. However, when measuring individual outcomes is costly, as is the case of a tumor biopsy, a sample-efficient strategy for acquiring each result is required. Deep Bayesian active learning provides a framework for efficient data acquisition by selecting points with high uncertainty. However, existing methods bias training data acquisition towards regions of non-overlapping support between the treated and control populations. These are not sample-efficient because the treatment effect is not identifiable in such regions. We introduce causal, Bayesian acquisition functions grounded in information theory that bias data acquisition towards regions with overlapping support to maximize sample efficiency for learning personalized treatment effects. We demonstrate the performance of the proposed acquisition strategies on synthetic and semi-synthetic datasets IHDP and CMNIST and their extensions, which aim to simulate common dataset biases and pathologies.
****
Scalable Rule-Based Representation Learning for Interpretable Classification
Zhuo Wang, Wei Zhang, Ning Liu, Jianyong Wang
Rule-based models, e.g., decision trees, are widely used in scenarios demanding high model interpretability for their transparent inner structures and good model expressivity. However, rule-based models are hard to optimize, especially on large data sets, due to their discrete parameters and structures. Ensemble methods and fuzzy/soft rules are commonly used to improve performance, but they sacrifice the model interpretability. To obtain both good scalability and interpretability, we propose a new classifier, named Rule-based Representation Learner (RRL)

, that automatically learns interpretable non-fuzzy rules for data representatio
n and classification. To train the non-differentiable RRL effectively, we projec
t it to a continuous space and propose a novel training method, called Gradient
Grafting, that can directly optimize the discrete model using gradient descent.
An improved design of logical activation functions is also devised to increase t
he scalability of RRL and enable it to discretize the continuous features end-to
-end. Exhaustive experiments on nine small and four large data sets show that RR
L outperforms the competitive interpretable approaches and can be easily adjuste
d to obtain a trade-off between classification accuracy and model complexity for
 different scenarios. Our code is available at: https://github.com/12wang3/rrl.
****
Bridging Non Co-occurrence with Unlabeled In-the-wild Data for Incremental Objec
t Detection
NA DONG, Yongqiang Zhang, Mingli Ding, Gim Hee Lee
Deep networks have shown remarkable results in the task of object detection. How
ever, their performance suffers critical drops when they are subsequently traine
d on novel classes without any sample from the base classes originally used to t
rain the model. This phenomenon is known as catastrophic forgetting. Recently, s
everal incremental learning methods are proposed to mitigate catastrophic forget
ting for object detection. Despite the effectiveness, these methods require co-o
ccurrence of the unlabeled base classes in the training data of the novel classe
s. This requirement is impractical in many real-world settings since the base cl
asses do not necessarily co-occur with the novel classes. In view of this limita
tion, we consider a more practical setting of complete absence of co-occurrence
of the base and novel classes for the object detection task. We propose the use
of unlabeled in-the-wild data to bridge the non co-occurrence caused by the miss
ing base classes during the training of additional novel classes. To this end, w
e introduce a blind sampling strategy based on the responses of the base-class m
odel and pre-trained novel-class model to select a smaller relevant dataset from
 the large in-the-wild dataset for incremental learning. We then design a dual-t
eacher distillation framework to transfer the knowledge distilled from the base-
 and novel-class teacher models to the student model using the sampled in-the-wi
ld data. Experimental results on the PASCAL VOC and MS COCO datasets show that o
ur proposed method significantly outperforms other state-of-the-art class-increm
ental object detection methods when there is no co-occurrence between the base a
nd novel classes during training.
****
A Regression Approach to Learning-Augmented Online Algorithms
Keerti Anand, Rong Ge, Amit Kumar, Debmalya Panigrahi
The emerging field of learning-augmented online algorithms uses ML techniques to
 predict future input parameters and thereby improve the performance of online a
lgorithms. Since these parameters are, in general, real-valued functions, a natu
ral approach is to use regression techniques to make these predictions. We intro
duce this approach in this paper, and explore it in the context of a general onl
ine search framework that captures classic problems like (generalized) ski renta
l, bin packing, minimum makespan scheduling, etc. We show nearly tight bounds on
 the sample complexity of this regression problem, and extend our results to the
 agnostic setting. From a technical standpoint, we show that the key is to incor
porate online optimization benchmarks in the design of the loss function for the
 regression problem, thereby diverging from the use of off-the-shelf regression
tools with standard bounds on statistical error.
****