Sign in to GitHub · GitHub
****
Data Structures for Density Estimation
Anders Aamand, Alexandr Andoni, Justin Y. Chen, Piotr Indyk, Shyam Narayanan, Sa ndeep Silwal
We study statistical/computational tradeoffs for the following density estimatio n problem: given k distributions $v_1, \ldots, v_k$ over a discrete domain of size n, and sampling access to a distribution p, identify v_i that is "clo se" to p. Our main result is the first data structure that, given a sublinear (in n) number of samples from p, identifies v_i in time sublinear in k. We also give an improved version of the algorithm of Acharya et al. (2018) that reports v_i in time linear in k. The experimental evaluation of the latter a lgorithm shows that it achieves a significant reduction in the number of operati ons needed to achieve a given accuracy compared to prior work.
****
ClusterFuG: Clustering Fully connected Graphs by Multicut
Ahmed Abbas, Paul Swoboda
We propose a graph clustering formulation based on multicut (a.k.a. weighted cor relation clustering) on the complete graph. Our formulation does not need specif ication of the graph topology as in the original sparse formulation of multicut, making our approach simpler and potentially better performing. In contrast to u nweighted correlation clustering we allow for a more expressive weighted cost st ructure. In dense multicut, the clustering objective is given in a factorized fo rm as inner products of node feature vectors. This allows for an efficient formu lation and inference in contrast to multicut/weighted correlation clustering, wh ich has at least quadratic representation and computation complexity when workin g on the complete graph. We show how to rewrite classical greedy algorithms for multicut in our dense setting and how to modify them for greater efficiency and solution quality. In particular, our algorithms scale to graphs with tens of tho usands of nodes. Empirical evidence on instance segmentation on Cityscapes and c lustering of ImageNet datasets shows the merits of our approach.
****
Generalization on the Unseen, Logic Reasoning and Degree Curriculum
Emmanuel Abbe, Samy Bengio, Aryo Lotfi, Kevin Rizk
This paper considers the learning of logical (Boolean) functions with focus on t he generalization on the unseen (GOTU) setting, a strong case of out-of-distribu tion generalization. This is motivated by the fact that the rich combinatorial n ature of data in certain reasoning tasks (e.g., arithmetic/logic) makes represen tative data sampling challenging, and learning successfully under GOTU gives a f irst vignette of an ’extrapolating’ or ’reasoning’ learner. We then study how di fferent network architectures trained by (S)GD perform under GOTU and provide bo th theoretical and experimental evidence that for a class of network models incl uding instances of Transformers, random features models, and diagonal linear net works, a min-degree-interpolator is learned on the unseen. We also provide evide nce that other instances with larger learning rates or mean-field networks reach leaky min-degree solutions. These findings lead to two implications: (1) we pro vide an explanation to the length generalization problem (e.g., Anil et al. 2022); (2) we introduce a curriculum learning algorithm called Degree-Curriculum tha t learns monomials more efficiently by incrementing supports.
****
Toward Large Kernel Models
Amirhesam Abedsoltan, Mikhail Belkin, Parthe Pandit
Recent studies indicate that kernel machines can often perform similarly or bett er than deep neural networks (DNNs) on small datasets. The interest in kernel ma chines has been additionally bolstered by the discovery of their equivalence to wide neural networks in certain regimes. However, a key feature of DNNs is their ability to scale the model size and training data size independently, whereas i n traditional kernel machines model size is tied to data size. Because of this c oupling, scaling kernel machines to large data has been computationally challeng ing. In this paper, we provide a way forward for constructing large-scale genera

l kernel models, which are a generalization of kernel machines that decouples th
e model and data, allowing training on large datasets. Specifically, we introduc
e EigenPro 3.0, an algorithm based on projected dual preconditioned SGD and show
 scaling to model and data sizes which have not been possible with existing kern
el methods. We provide a PyTorch based implementation which can take advantage o
f multiple GPUs.
****

Expertise Trees Resolve Knowledge Limitations in Collective Decision-Making
Axel Abels, Tom Lenaerts, Vito Trianni, Ann Nowe
Experts advising decision-makers are likely to display expertise which varies as
 a function of the problem instance. In practice, this may lead to sub-optimal o
r discriminatory decisions against minority cases. In this work, we model such c
hanges in depth and breadth of knowledge as a partitioning of the problem space
into regions of differing expertise. We provide here new algorithms that explici
tly consider and adapt to the relationship between problem instances and experts
’ knowledge. We first propose and highlight the drawbacks of a naive approach ba
sed on nearest neighbor queries. To address these drawbacks we then introduce a
novel algorithm — expertise trees — that constructs decision trees enabling the
learner to select appropriate models. We provide theoretical insights and empiri
cally validate the improved performance of our novel approach on a range of prob
lems for which existing methods proved to be inadequate.
****

Comparison of meta-learners for estimating multi-valued treatment heterogeneous
effects
Naoufal Acharki, Ramiro Lugo, Antoine Bertoncello, Josselin Garnier
Conditional Average Treatment Effects (CATE) estimation is one of the main chall
enges in causal inference with observational data. In addition to Machine Learni
ng based-models, nonparametric estimators called meta-learners have been develop
ed to estimate the CATE with the main advantage of not restraining the estimatio
n to a specific supervised learning method. This task becomes, however, more com
plicated when the treatment is not binary as some limitations of the naive exten
sions emerge. This paper looks into meta-learners for estimating the heterogeneo
us effects of multi-valued treatments. We consider different meta-learners, and
we carry out a theoretical analysis of their error upper bounds as functions of
important parameters such as the number of treatment levels, showing that the na
ive extensions do not always provide satisfactory results. We introduce and disc
uss meta-learners that perform well as the number of treatments increases. We em
pirically confirm the strengths and weaknesses of those methods with synthetic a
nd semi-synthetic datasets.
****

BNN-DP: Robustness Certification of Bayesian Neural Networks via Dynamic Program
ming
Steven Adams, Andrea Patane, Morteza Lahijanian, Luca Laurenti
In this paper, we introduce BNN-DP, an efficient algorithmic framework for analy
sis of adversarial robustness of Bayesian Neural Networks (BNNs). Given a compac
t set of input points $T\subset \mathbb{R}^n$, BNN-DP computes lower and upper b
ounds on the BNN’s predictions for all the points in T. The framework is based
 on an interpretation of BNNs as stochastic dynamical systems, which enables the
 use of Dynamic Programming (DP) algorithms to bound the prediction range along
the layers of the network. Specifically, the method uses bound propagation techn
iques and convex relaxations to derive a backward recursion procedure to over-ap
proximate the prediction range of the BNN with piecewise affine functions. The a
lgorithm is general and can handle both regression and classification tasks. On
a set of experiments on various regression and classification tasks and BNN arch
itectures, we show that BNN-DP outperforms state-of-the-art methods by up to fou
r orders of magnitude in both tightness of the bounds and computational efficien
cy.
****

SAM operates far from home: eigenvalue regularization as a dynamical phenomenon
Atish Agarwala, Yann Dauphin

The Sharpness Aware Minimization (SAM) optimization algorithm has been shown to control large eigenvalues of the loss Hessian and provide generalization benefits in a variety of settings. The original motivation for SAM was a modified loss function which penalized sharp minima; subsequent analyses have also focused on the behavior near minima. However, our work reveals that SAM provides a strong regularization of the eigenvalues throughout the learning trajectory. We show that in a simplified setting, SAM dynamically induces a stabilization related to the edge of stability (EOS) phenomenon observed in large learning rate gradient descent. Our theory predicts the largest eigenvalue as a function of the learning rate and SAM radius parameters. Finally, we show that practical models can also exhibit this EOS stabilization, and that understanding SAM must account for these dynamics far away from any minima.

****

Second-order regression models exhibit progressive sharpening to the edge of stability

Atish Agarwala, Fabian Pedregosa, Jeffrey Pennington

Recent studies of gradient descent with large step sizes have shown that there is often a regime with an initial increase in the largest eigenvalue of the loss Hessian (progressive sharpening), followed by a stabilization of the eigenvalue near the maximum value which allows convergence (edge of stability). These phenomena are intrinsically non-linear and do not happen for models in the constant Neural Tangent Kernel (NTK) regime, for which the predictive function is approximately linear in the parameters. As such, we consider the next simplest class of predictive models, namely those that are quadratic in the parameters, which we call second-order regression models. For quadratic objectives in two dimensions, we prove that this second-order regression model exhibits progressive sharpening of the NTK eigenvalue towards a value that differs slightly from the edge of stability, which we explicitly compute. In higher dimensions, the model generically shows similar behavior, even without the specific structure of a neural network, suggesting that progressive sharpening and edge-of-stability behavior aren’t unique features of neural networks, and could be a more general property of discrete learning algorithms in high-dimensional non-linear models.

****

Global optimality of Elman-type RNNs in the mean-field regime

Andrea Agazzi, Jianfeng Lu, Sayan Mukherjee

We analyze Elman-type recurrent neural networks (RNNs) and their training in the mean-field regime. Specifically, we show convergence of gradient descent training dynamics of the RNN to the corresponding mean-field formulation in the large width limit. We also show that the fixed points of the limiting infinite-width dynamics are globally optimal, under some assumptions on the initialization of the weights. Our results establish optimality for feature-learning with wide RNNs in the mean-field regime.

****

SemSup-XC: Semantic Supervision for Zero and Few-shot Extreme Classification

Pranjal Aggarwal, Ameet Deshpande, Karthik R Narasimhan

Extreme classification (XC) involves predicting over large numbers of classes (thousands to millions), with real-world applications like news article classification and e-commerce product tagging. The zero-shot version of this task requires generalization to novel classes without additional supervision. In this paper, we develop SemSup-XC, a model that achieves state-of-the-art zero-shot and few-shot performance on three XC datasets derived from legal, e-commerce, and Wikipedia data. To develop SemSup-XC, we use automatically collected semantic class descriptions to represent classes and facilitate generalization through a novel hybrid matching module that matches input instances to class descriptions using a combination of semantic and lexical similarity. Trained with contrastive learning, SemSup-XC significantly outperforms baselines and establishes state-of-the-art performance on all three datasets considered, gaining up to 12 precision points on zero-shot and more than 10 precision points on one-shot tests, with similar gains for recall@10. Our ablation studies highlight the relative importance of our hybrid matching module and automatically collected class descriptions.

****
Adaptive IMLE for Few-shot Pretraining-free Generative Modelling
Mehran Aghabozorgi, Shichong Peng, Ke Li
Despite their success on large datasets, GANs have been difficult to apply in th e few-shot setting, where only a limited number of training examples are provide d. Due to mode collapse, GANs tend to ignore some training examples, causing ove rfitting to a subset of the training dataset, which is small in the first place. A recent method called Implicit Maximum Likelihood Estimation (IMLE) is an alte rnative to GAN that tries to address this issue. It uses the same kind of genera tors as GANs but trains it with a different objective that encourages mode cover age. However, the theoretical guarantees of IMLE hold under a restrictive condit ion that the optimal likelihood at all data points is the same. In this paper, w e present a more generalized formulation of IMLE which includes the original for mulation as a special case, and we prove that the theoretical guarantees hold un der weaker conditions. Using this generalized formulation, we further derive a n ew algorithm, which we dub Adaptive IMLE, which can adapt to the varying difficu lty of different training examples. We demonstrate on multiple few-shot image sy nthesis datasets that our method significantly outperforms existing methods. Our code is available at https://github.com/mehranagh20/AdaIMLE.
****
Scaling Laws for Generative Mixed-Modal Language Models
Armen Aghajanyan, Lili Yu, Alexis Conneau, Wei-Ning Hsu, Karen Hambardzumyan, Su san Zhang, Stephen Roller, Naman Goyal, Omer Levy, Luke Zettlemoyer
Generative language models define distributions over sequences of tokens that ca n represent essentially any combination of data modalities (e.g., any permutatio n of image tokens from VQ-VAEs, speech tokens from HuBERT, BPE tokens for langua ge or code, and so on). To better understand the scaling properties of such mixe d-modal models, we conducted over 250 experiments using seven different modaliti es and model sizes ranging from 8 million to 30 billion, trained on 5-100 billio n tokens. We report new mixed-modal scaling laws that unify the contributions of individual modalities and the interactions between them. Specifically, we expli citly model the optimal synergy and competition due to data and model size as an additive term to previous uni-modal scaling laws. We also find four empirical p henomena observed during the training, such as emergent coordinate-ascent style training that naturally alternates between modalities, guidelines for selecting critical hyper-parameters, and connections between mixed-modal competition and t raining stability. Finally, we test our scaling law by training a 30B speech-tex t model, which significantly outperforms the corresponding unimodal models. Over all, our research provides valuable insights into the design and training of mix ed-modal generative models, an important new class of unified models that have u nique distributional properties.
****
Hypothesis Transfer Learning with Surrogate Classification Losses: Generalizatio n Bounds through Algorithmic Stability
Anass Aghbalou, Guillaume Staerman
Hypothesis transfer learning (HTL) contrasts domain adaptation by allowing for a previous task leverage, named the source, into a new one, the target, without r equiring access to the source data. Indeed, HTL relies only on a hypothesis lear nt from such source data, relieving the hurdle of expansive data storage and pro viding great practical benefits. Hence, HTL is highly beneficial for real-world applications relying on big data. The analysis of such a method from a theoretic al perspective faces multiple challenges, particularly in classification tasks. This paper deals with this problem by studying the learning theory of HTL throug h algorithmic stability, an attractive theoretical framework for machine learnin g algorithms analysis. In particular, we are interested in the statistical behav ior of the regularized empirical risk minimizers in the case of binary classific ation. Our stability analysis provides learning guarantees under mild assumption s. Consequently, we derive several complexity-free generalization bounds for ess ential statistical quantities like the training error, the excess risk and cross -validation estimates. These refined bounds allow understanding the benefits of

transfer learning and comparing the behavior of standard losses in different sce
narios, leading to valuable insights for practitioners.
****

Constrained Causal Bayesian Optimization
Virginia Aglietti, Alan Malek, Ira Ktena, Silvia Chiappa
We propose constrained causal Bayesian optimization (cCBO), an approach for find
ing interventions in a known causal graph that optimize a target variable under
some constraints. cCBO first reduces the search space by exploiting the graph st
ructure and, if available, an observational dataset; and then solves the restric
ted optimization problem by modelling target and constraint quantities using Gau
ssian processes and by sequentially selecting interventions via a constrained ex
pected improvement acquisition function. We propose different surrogate models t
hat enable to integrate observational and interventional data while capturing co
rrelation among effects with increasing levels of sophistication. We evaluate cC
BO on artificial and real-world causal graphs showing successful trade off betwe
en fast convergence and percentage of feasible interventions.
****

Explaining the effects of non-convergent MCMC in the training of Energy-Based Mo
dels
Elisabeth Agoritsas, Giovanni Catania, Aurélien Decelle, Beatriz Seoane
In this paper, we quantify the impact of using non-convergent Markov chains to t
rain Energy-Based models (EBMs). In particular, we show analytically that EBMs t
rained with non-persistent short runs to estimate the gradient can perfectly rep
roduce a set of empirical statistics of the data, not at the level of the equili
brium measure, but through a precise dynamical process. Our results provide a fi
rst-principles explanation for the observations of recent works proposing the st
rategy of using short runs starting from random initial conditions as an efficie
nt way to generate high-quality samples in EBMs, and lay the groundwork for usin
g EBMs as diffusion models. After explaining this effect in generic EBMs, we ana
lyze two solvable models in which the effect of the non-convergent sampling in t
he trained parameters can be described in detail. Finally, we test these predict
ions numerically on a ConvNet EBM and a Boltzmann machine.
****

Using Large Language Models to Simulate Multiple Humans and Replicate Human Subj
ect Studies
Gati V Aher, Rosa I. Arriaga, Adam Tauman Kalai
We introduce a new type of test, called a Turing Experiment (TE), for evaluating
 to what extent a given language model, such as GPT models, can simulate differe
nt aspects of human behavior. A TE can also reveal consistent distortions in a l
anguage model’s simulation of a specific human behavior. Unlike the Turing Test,
 which involves simulating a single arbitrary individual, a TE requires simulati
ng a representative sample of participants in human subject research. We carry o
ut TEs that attempt to replicate well-established findings from prior studies. W
e design a methodology for simulating TEs and illustrate its use to compare how
well different language models are able to reproduce classic economic, psycholin
guistic, and social psychology experiments: Ultimatum Game, Garden Path Sentence
s, Milgram Shock Experiment, and Wisdom of Crowds. In the first three TEs, the e
xisting findings were replicated using recent models, while the last TE reveals
a “hyper-accuracy distortion” present in some language models (including ChatGPT
 and GPT-4), which could affect downstream applications in education and the art
s.
****

Interventional Causal Representation Learning
Kartik Ahuja, Divyat Mahajan, Yixin Wang, Yoshua Bengio
Causal representation learning seeks to extract high-level latent factors from l
ow-level sensory data. Most existing methods rely on observational data and stru
ctural assumptions (e.g., conditional independence) to identify the latent facto
rs. However, interventional data is prevalent across applications. Can intervent
ional data facilitate causal representation learning? We explore this question i
n this paper. The key observation is that interventional data often carries geom

etric signatures of the latent factors’ support (i.e. what values each latent ca n possibly take). For example, when the latent factors are causally connected, i nterventions can break the dependency between the intervened latents’ support an d their ancestors’. Leveraging this fact, we prove that the latent causal factor s can be identified up to permutation and scaling given data from perfect do int erventions. Moreover, we can achieve block affine identification, namely the est imated latent factors are only entangled with a few other latents if we have acc ess to data from imperfect interventions. These results highlight the unique pow er of interventional data in causal representation learning; they can enable pro vable identification of latent factors without any assumptions about their distr ibutions or dependency structure.

****

Sequential Underspecified Instrument Selection for Cause-Effect Estimation
Elisabeth Ailer, Jason Hartford, Niki Kilbertus

Instrumental variable (IV) methods are used to estimate causal effects in settin gs with unobserved confounding, where we cannot directly experiment on the treat ment variable. Instruments are variables which only affect the outcome indirectl y via the treatment variable(s). Most IV applications focus on low-dimensional t reatments and crucially require at least as many instruments as treatments. This assumption is restrictive: in the natural sciences we often seek to infer causa l effects of high-dimensional treatments (e.g., the effect of gene expressions o r microbiota on health and disease), but can only run few experiments with a lim ited number of instruments (e.g., drugs or antibiotics). In such under-specified problems, the full treatment effect is not identifiable in a single experiment even in the linear case. We show that one can still reliably recover the project ion of the treatment effect onto the instrumented subspace and develop technique s to consistently combine such partial estimates from different sets of instrume nts. We then leverage our combined estimators in an algorithm that iteratively p roposes the most informative instruments at each round of experimentation to max imize the overall information about the full causal effect.

****

Atari-5: Distilling the Arcade Learning Environment down to Five Games
Matthew Aitchison, Penny Sweetser, Marcus Hutter

The Arcade Learning Environment (ALE) has become an essential benchmark for asse ssing the performance of reinforcement learning algorithms. However, the computa tional cost of generating results on the entire 57-game dataset limits ALE’s use and makes the reproducibility of many results infeasible. We propose a novel so lution to this problem in the form of a principled methodology for selecting sma ll but representative subsets of environments within a benchmark suite. We appli ed our method to identify a subset of five ALE games, we call Atari-5, which pro duces 57-game median score estimates within 10% of their true values. Extending the subset to 10-games recovers 80% of the variance for log-scores for all games within the 57-game set. We show this level of compression is possible due to a high degree of correlation between many of the games in ALE.

****

Towards credible visual model interpretation with path attribution
Naveed Akhtar, Mohammad A. A. K. Jalwana

With its inspirational roots in game-theory, path attribution framework stands o ut among the post-hoc model interpretation techniques due to its axiomatic natur e. However, recent developments show that despite being axiomatic, path attribut ion methods can compute counter-intuitive feature attributions. Not only that, f or deep visual models, the methods may also not conform to the original game-the oretic intuitions that are the basis of their axiomatic nature. To address these issues, we perform a systematic investigation of the path attribution framework . We first pinpoint the conditions in which the counter-intuitive attributions o f deep visual models can be avoided under this framework. Then, we identify a me chanism of integrating the attributions over the paths such that they computatio nally conform to the original insights of game-theory. These insights are eventu ally combined into a method, which provides intuitive and reliable feature attri butions. We also establish the findings empirically by evaluating the method on

multiple datasets, models and evaluation metrics. Extensive experiments show a c
onsistent quantitative and qualitative gain in the results over the baselines.
****
Convergence of First-Order Methods for Constrained Nonconvex Optimization with D
ependent Data
Ahmet Alacaoglu, Hanbaek Lyu
We focus on analyzing the classical stochastic projected gradient methods under
a general dependent data sampling scheme for constrained smooth nonconvex optimi
zation. We show the worst-case rate of convergence $\tilde{O}(t^{-1/4})$ and com
plexity $\tilde{O}(\varepsilon^{-4})$ for achieving an ε-near statio
nary point in terms of the norm of the gradient of Moreau envelope and gradient
mapping. While classical convergence guarantee requires i.i.d. data sampling fro
m the target distribution, we only require a mild mixing condition of the condit
ional distribution, which holds for a wide class of Markov chain sampling algori
thms. This improves the existing complexity for the constrained smooth nonconvex
optimization with dependent data from $\tilde{O}(\varepsilon^{-8})$ to $\tilde{
O}(\varepsilon^{-4})$ with a significantly simpler analysis. We illustrate the g
enerality of our approach by deriving convergence results with dependent data fo
r stochastic proximal gradient methods, adaptive stochastic gradient algorithm A
daGrad and stochastic gradient algorithm with heavy ball momentum. As an applica
tion, we obtain first online nonnegative matrix factorization algorithms for dep
endent data based on stochastic projected gradient methods with adaptive step si
zes and optimal rate of convergence.
****
Recasting Self-Attention with Holographic Reduced Representations
Mohammad Mahmudul Alam, Edward Raff, Stella Biderman, Tim Oates, James Holt
In recent years, self-attention has become the dominant paradigm for sequence mo
deling in a variety of domains. However, in domains with very long sequence leng
ths the $\mathcal{O}(T^2)$ memory and $\mathcal{O}(T^2 H)$ compute costs can mak
e using transformers infeasible. Motivated by problems in malware detection, whe
re sequence lengths of $T \geq 100,000$ are a roadblock to deep learning, we re-
cast self-attention using the neuro-symbolic approach of Holographic Reduced Rep
resentations (HRR). In doing so we perform the same high-level strategy of the s
tandard self-attention: a set of queries matching against a set of keys, and ret
urning a weighted response of the values for each key. Implemented as a “Hrrform
er” we obtain several benefits including $\mathcal{O}(T H \log H)$ time complexi
ty, $\mathcal{O}(T H)$ space complexity, and convergence in $10\times$ fewer epo
chs. Nevertheless, the Hrrformer achieves near state-of-the-art accuracy on LRA
benchmarks and we are able to learn with just a single layer. Combined, these be
nefits make our Hrrformer the first viable Transformer for such long malware cla
ssification sequences and up to $280\times$ faster to train on the Long Range Ar
ena benchmark.
****
The Saddle-Point Method in Differential Privacy
Wael Alghamdi, Juan Felipe Gomez, Shahab Asoodeh, Flavio Calmon, Oliver Kosut, L
alitha Sankar
We characterize the differential privacy guarantees of privacy mechanisms in the
large-composition regime, i.e., when a privacy mechanism is sequentially applie
d a large number of times to sensitive data. Via exponentially tilting the priva
cy loss random variable, we derive a new formula for the privacy curve expressin
g it as a contour integral over an integration path that runs parallel to the im
aginary axis with a free real-axis intercept. Then, using the method of steepest
descent from mathematical physics, we demonstrate that the choice of saddle-poi
nt as the real-axis intercept yields closed-form accurate approximations of the
desired contour integral. This procedure—dubbed the saddle-point accountant (SPA
)—yields a constant-time accurate approximation of the privacy curve. Theoretica
lly, our results can be viewed as a refinement of both Gaussian Differential Pri
vacy and the moments accountant method found in Rényi Differential Privacy. In p
ractice, we demonstrate through numerical experiments that the SPA provides a pr
ecise approximation of privacy guarantees competitive with purely numerical-base

d methods (such as FFT-based accountants), while enjoying closed-form mathematic al expressions.
****
Nonlinear Advantage: Trained Networks Might Not Be As Complex as You Think
Christian H.X. Ali Mehmeti-Göpel, Jan Disselhoff
We perform an empirical study of the behaviour of deep networks when fully linea rizing some of its feature channels through a sparsity prior on the overall numb er of nonlinear units in the network. In experiments on image classification and machine translation tasks, we investigate how much we can simplify the network function towards linearity before performance collapses. First, we observe a sig nificant performance gap when reducing nonlinearity in the network function earl y on as opposed to late in training, in-line with recent observations on the tim e-evolution of the data-dependent NTK. Second, we find that after training, we a re able to linearize a significant number of nonlinear units while maintaining a high performance, indicating that much of a network’s expressivity remains unus ed but helps gradient descent in early stages of training. To characterize the d epth of the resulting partially linearized network, we introduce a measure calle d average path length, representing the average number of active nonlinearities encountered along a path in the network graph. Under sparsity pressure, we find that the remaining nonlinear units organize into distinct structures, forming co re-networks of near constant effective depth and width, which in turn depend on task difficulty.
****
A Simple Zero-shot Prompt Weighting Technique to Improve Prompt Ensembling in Te xt-Image Models
James Urquhart Allingham, Jie Ren, Michael W Dusenberry, Xiuye Gu, Yin Cui, Dust in Tran, Jeremiah Zhe Liu, Balaji Lakshminarayanan
Contrastively trained text-image models have the remarkable ability to perform z ero-shot classification, that is, classifying previously unseen images into cate gories that the model has never been explicitly trained to identify. However, th ese zero-shot classifiers need prompt engineering to achieve high accuracy. Prom pt engineering typically requires hand-crafting a set of prompts for individual downstream tasks. In this work, we aim to automate this prompt engineering and i mprove zero-shot accuracy through prompt ensembling. In particular, we ask “Give n a large pool of prompts, can we automatically score the prompts and ensemble t hose that are most suitable for a particular downstream dataset, without needing access to labeled validation data?". We demonstrate that this is possible. In d oing so, we identify several pathologies in a naive prompt scoring method where the score can be easily overconfident due to biases in pre-training and test dat a, and we propose a novel prompt scoring method that corrects for the biases. Us ing our proposed scoring method to create a weighted average prompt ensemble, ou r method overall outperforms equal average ensemble, as well as hand-crafted pro mpts, on ImageNet, 4 of its variants, and 11 fine-grained classification benchma rks. while being fully automatic, optimization-free, and not requiring access to labeled validation data.
****
On the Privacy-Robustness-Utility Trilemma in Distributed Learning
Youssef Allouah, Rachid Guerraoui, Nirupam Gupta, Rafael Pinot, John Stephan
The ubiquity of distributed machine learning (ML) in sensitive public domain app lications calls for algorithms that protect data privacy, while being robust to faults and adversarial behaviors. Although privacy and robustness have been exte nsively studied independently in distributed ML, their synthesis remains poorly understood. We present the first tight analysis of the error incurred by any alg orithm ensuring robustness against a fraction of adversarial machines, as well a s differential privacy (DP) for honest machines’ data against any other curious entity. Our analysis exhibits a fundamental trade-off between privacy, robustnes s, and utility. To prove our lower bound, we consider the case of mean estimatio n, subject to distributed DP and robustness constraints, and devise reductions t o centralized estimation of one-way marginals. We prove our matching upper bound by presenting a new distributed ML algorithm using a high-dimensional robust ag

gregation rule. The latter amortizes the dependence on the dimension in the erro
r (caused by adversarial workers and DP), while being agnostic to the statistica
l properties of the data.
****

Differentially Private Distributed Bayesian Linear Regression with MCMC
Baris Alparslan, Sinan Y■ld■r■m, Ilker Birbil
We propose a novel Bayesian inference framework for distributed differentially p
rivate linear regression. We consider a distributed setting where multiple parti
es hold parts of the data and share certain summary statistics of their portions
 in privacy-preserving noise. We develop a novel generative statistical model fo
r privately shared statistics, which exploits a useful distributional relation b
etween the summary statistics of linear regression. We propose Bayesian estimati
on of the regression coefficients, mainly using Markov chain Monte Carlo algorit
hms, while we also provide a fast version that performs approximate Bayesian est
imation in one iteration. The proposed methods have computational advantages ove
r their competitors. We provide numerical results on both real and simulated dat
a, which demonstrate that the proposed algorithms provide well-rounded estimatio
n and prediction.
****

Robust and Scalable Bayesian Online Changepoint Detection
Matias Altamirano, Francois-Xavier Briol, Jeremias Knoblauch
This paper proposes an online, provably robust, and scalable Bayesian approach f
or changepoint detection. The resulting algorithm has key advantages over previo
us work: it provides provable robustness by leveraging the generalised Bayesian
perspective, and also addresses the scalability issues of previous attempts. Spe
cifically, the proposed generalised Bayesian formalism leads to conjugate poster
iors whose parameters are available in closed form by leveraging diffusion score
 matching. The resulting algorithm is exact, can be updated through simple algeb
ra, and is more than 10 times faster than its closest competitor.
****

Neural Wasserstein Gradient Flows for Discrepancies with Riesz Kernels
Fabian Altekrüger, Johannes Hertrich, Gabriele Steidl
Wasserstein gradient flows of maximum mean discrepancy (MMD) functionals with no
n-smooth Riesz kernels show a rich structure as singular measures can become abs
olutely continuous ones and conversely. In this paper we contribute to the under
standing of such flows. We propose to approximate the backward scheme of Jordan,
 Kinderlehrer and Otto for computing such Wasserstein gradient flows as well as
a forward scheme for so-called Wasserstein steepest descent flows by neural netw
orks (NNs). Since we cannot restrict ourselves to absolutely continuous measures
, we have to deal with transport plans and velocity plans instead of usual trans
port maps and velocity fields. Indeed, we approximate the disintegration of both
 plans by generative NNs which are learned with respect to appropriate loss func
tions. In order to evaluate the quality of both neural schemes, we benchmark the
m on the interaction energy. Here we provide analytic formulas for Wasserstein s
chemes starting at a Dirac measure and show their convergence as the time step s
ize tends to zero. Finally, we illustrate our neural MMD flows by numerical exam
ples.
****

Distributed Contextual Linear Bandits with Minimax Optimal Communication Cost
Sanae Amani, Tor Lattimore, András György, Lin Yang
We study distributed contextual linear bandits with stochastic contexts, where $
N$ agents/learners act cooperatively to solve a linear bandit-optimization probl
em with d-dimensional features over the course of T rounds. For this problem
, we derive the first ever information-theoretic lower bound $\Omega(dN)$ on the
 communication cost of any algorithm that performs optimally in a regret minimiz
ation setup. We then propose a distributed batch elimination version of the LinU
CB algorithm, DisBE-LUCB, where the agents share information among each other th
rough a central server. We prove that the communication cost of DisBE-LUCB, matc
hes our lower bound up to logarithmic factors. In particular, for scenarios with
 known context distribution, the communication cost of DisBE-LUCB is only $\tild

e{\mathcal{O}}(dN)$ and its regret is $\tilde{\mathcal{O}}(\sqrt{dNT})$, which i
s of the same order as that incurred by an optimal single-agent algorithm for $N
T$ rounds. We also provide similar bounds for practical settings where the conte
xt distribution can only be estimated. Therefore, our proposed algorithm is near
ly minimax optimal in terms of both regret and communication cost. Finally, we p
ropose DecBE-LUCB, a fully decentralized version of DisBE-LUCB, which operates w
ithout a central server, where agents share information with their immediate nei
ghbors through a carefully designed consensus procedure.
****

A Kernelized Stein Discrepancy for Biological Sequences
Alan Nawzad Amin, Eli N Weinstein, Debora Susan Marks
Generative models of biological sequences are a powerful tool for learning from
complex sequence data, predicting the effects of mutations, and designing novel
biomolecules with desired properties. To evaluate generative models it is import
ant to accurately measure differences between high-dimensional distributions. In
this paper we propose the “KSD-B”, a novel divergence measure for distributions
over biological sequences that is based on the kernelized Stein discrepancy (KS
D). The KSD-B can be evaluated even when the normalizing constant of the model i
s unknown; it allows for variable length sequences and can take into account bio
logical notions of sequence distance. Unlike previous KSDs over discrete spaces
the KSD-B (a) is theoretically guaranteed to detect convergence and non-converge
nce of distributions over sequence space and (b) can be efficiently estimated in
practice. We demonstrate the advantages of the KSD-B on problems with synthetic
and real data, and apply it to measure the fit of state-of-the-art machine lear
ning models. Overall, the KSD-B enables rigorous evaluation of generative biolog
ical sequence models, allowing the accuracy of models, sampling procedures, and
library designs to be checked reliably.
****

The Optimal Approximation Factors in Misspecified Off-Policy Value Function Esti
mation
Philip Amortila, Nan Jiang, Csaba Szepesvari
Theoretical guarantees in reinforcement learning (RL) are known to suffer multip
licative blow-up factors with respect to the misspecification error of function
approximation. Yet, the nature of such approximation factors—especially their op
timal form in a given learning problem—is poorly understood. In this paper we st
udy this question in linear off-policy value function estimation, where many ope
n questions remain. We study the approximation factor in a broad spectrum of set
tings, such as presence vs. absence of state aliasing and full vs. partial cover
age of the state space. Our core results include instance-dependent upper bounds
on the approximation factors with respect to both the weighted L_2-norm (wher
e the weighting is the offline state distribution) and the L_∞ norm. We s
how that these approximation factors are optimal (in an instance-dependent sense
) for a number of these settings. In other cases, we show that the instance-depe
ndent parameters which appear in the upper bounds are necessary, and that the fi
niteness of either alone cannot guarantee a finite approximation factor even in
the limit of infinite data.
****

Meta Optimal Transport
Brandon Amos, Giulia Luise, Samuel Cohen, Ievgen Redko
We study the use of amortized optimization to predict optimal transport (OT) map
s from the input measures, which we call Meta OT. This helps repeatedly solve si
milar OT problems between different measures by leveraging the knowledge and inf
ormation present from past problems to rapidly predict and solve new problems. O
therwise, standard methods ignore the knowledge of the past solutions and subopt
imally re-solve each problem from scratch. We instantiate Meta OT models in disc
rete and continuous settings between grayscale images, spherical data, classific
ation labels, and color palettes and use them to improve the computational time
of standard OT solvers. Our source code is available at http://github.com/facebo
okresearch/meta-ot
****

Near-Optimal $Φ$-Regret Learning in Extensive-Form Games
Ioannis Anagnostides, Gabriele Farina, Tuomas Sandholm
In this paper, we establish efficient and uncoupled learning dynamics so that, when employed by all players in multiplayer perfect-recall imperfect-information extensive-form games, the trigger regret of each player grows as $O(\log T)$ after T repetitions of play. This improves exponentially over the prior best known trigger-regret bound of $O(T^{1/4})$, and settles a recent open question by Bai et al. (2022). As an immediate consequence, we guarantee convergence to the set of extensive-form correlated equilibria and coarse correlated equilibria at a near-optimal rate of $\frac{\log T}{T}$. Building on prior work, at the heart of our construction lies a more general result regarding fixed points deriving from rational functions with polynomial degree, a property that we establish for the fixed points of (coarse) trigger deviation functions. Moreover, our construction leverages a refined regret circuit for the convex hull, which—unlike prior guarantees—preserves the RVU property introduced by Syrgkanis et al. (NIPS, 2015); this observation has an independent interest in establishing near-optimal regret under learning dynamics based on a CFR-type decomposition of the regret.
****
A Modern Look at the Relationship between Sharpness and Generalization
Maksym Andriushchenko, Francesco Croce, Maximilian Müller, Matthias Hein, Nicolas Flammarion
Sharpness of minima is a promising quantity that can correlate with generalization in deep networks and, when optimized during training, can improve generalization. However, standard sharpness is not invariant under reparametrizations of neural networks, and, to fix this, reparametrization-invariant sharpness definitions have been proposed, most prominently adaptive sharpness (Kwon et al., 2021). But does it really capture generalization in modern practical settings? We comprehensively explore this question in a detailed study of various definitions of adaptive sharpness in settings ranging from training from scratch on ImageNet and CIFAR-10 to fine-tuning CLIP on ImageNet and BERT on MNLI. We focus mostly on transformers for which little is known in terms of sharpness despite their widespread usage. Overall, we observe that sharpness does not correlate well with generalization but rather with some training parameters like the learning rate that can be positively or negatively correlated with generalization depending on the setup. Interestingly, in multiple cases, we observe a consistent negative correlation of sharpness with OOD generalization implying that sharper minima can generalize better. Finally, we illustrate on a simple model that the right sharpness measure is highly data-dependent, and that we do not understand well this aspect for realistic data distributions.
****
SGD with Large Step Sizes Learns Sparse Features
Maksym Andriushchenko, Aditya Vardhan Varre, Loucas Pillaud-Vivien, Nicolas Flammarion
We showcase important features of the dynamics of the Stochastic Gradient Descent (SGD) in the training of neural networks. We present empirical observations that commonly used large step sizes (i) may lead the iterates to jump from one side of a valley to the other causing loss stabilization, and (ii) this stabilization induces a hidden stochastic dynamics that biases it implicitly toward simple predictors. Furthermore, we show empirically that the longer large step sizes keep SGD high in the loss landscape valleys, the better the implicit regularization can operate and find sparse representations. Notably, no explicit regularization is used: the regularization effect comes solely from the SGD dynamics influenced by the large step sizes schedule. Therefore, these observations unveil how, through the step size schedules, both gradient and noise drive together the SGD dynamics through the loss landscape of neural networks. We justify these findings theoretically through the study of simple neural network models as well as qualitative arguments inspired from stochastic processes. This analysis allows us to shed new light on some common practices and observed phenomena when training deep networks.
****

Neural Continuous-Discrete State Space Models for Irregularly-Sampled Time Series
Abdul Fatir Ansari, Alvin Heng, Andre Lim, Harold Soh
Learning accurate predictive models of real-world dynamic phenomena (e.g., climate, biological) remains a challenging task. One key issue is that the data generated by both natural and artificial processes often comprise time series that are irregularly sampled and/or contain missing observations. In this work, we propose the Neural Continuous-Discrete State Space Model (NCDSSM) for continuous-time modeling of time series through discrete-time observations. NCDSSM employs auxiliary variables to disentangle recognition from dynamics, thus requiring amortized inference only for the auxiliary variables. Leveraging techniques from continuous-discrete filtering theory, we demonstrate how to perform accurate Bayesian inference for the dynamic states. We propose three flexible parameterizations of the latent dynamics and an efficient training objective that marginalizes the dynamic states during inference. Empirical results on multiple benchmark datasets across various domains show improved imputation and forecasting performance of NCDSSM over existing models.
****
Paging with Succinct Predictions
Antonios Antoniadis, Joan Boyar, Marek Elias, Lene Monrad Favrholdt, Ruben Hoeksma, Kim S. Larsen, Adam Polak, Bertrand Simon
Paging is a prototypical problem in the area of online algorithms. It has also played a central role in the development of learning-augmented algorithms. Previous work on learning-augmented paging has investigated predictions on (i) when the current page will be requested again (reoccurrence predictions), (ii) the current state of the cache in an optimal algorithm (state predictions), (iii) all requests until the current page gets requested again, and (iv) the relative order in which pages are requested. We study learning-augmented paging from the new perspective of requiring the least possible amount of predicted information. More specifically, the predictions obtained alongside each page request are limited to one bit only. We develop algorithms satisfy all three desirable properties of learning-augmented algorithms – that is, they are consistent, robust and smooth – despite being limited to a one-bit prediction per request. We also present lower bounds establishing that our algorithms are essentially best possible.
****
Mixing Predictions for Online Metric Algorithms
Antonios Antoniadis, Christian Coester, Marek Elias, Adam Polak, Bertrand Simon
A major technique in learning-augmented online algorithms is combining multiple algorithms or predictors. Since the performance of each predictor may vary over time, it is desirable to use not the single best predictor as a benchmark, but rather a dynamic combination which follows different predictors at different times. We design algorithms that combine predictions and are competitive against such dynamic combinations for a wide class of online problems, namely, metrical task systems. Against the best (in hindsight) unconstrained combination of ℓ predictors, we obtain a competitive ratio of $O(\ell^2)$, and show that this is best possible. However, for a benchmark with slightly constrained number of switches between different predictors, we can get a $(1+\epsilon)$-competitive algorithm. Moreover, our algorithms can be adapted to access predictors in a bandit-like fashion, querying only one predictor at a time. An unexpected implication of one of our lower bounds is a new structural insight about covering formulations for the k-server problem.
****
Exponential Smoothing for Off-Policy Learning
Imad Aouali, Victor-Emmanuel Brunel, David Rohde, Anna Korba
Off-policy learning (OPL) aims at finding improved policies from logged bandit data, often by minimizing the inverse propensity scoring (IPS) estimator of the risk. In this work, we investigate a smooth regularization for IPS, for which we derive a two-sided PAC-Bayes generalization bound. The bound is tractable, scalable, interpretable and provides learning certificates. In particular, it is also valid for standard IPS without making the assumption that the importance weight

s are bounded. We demonstrate the relevance of our approach and its favorable pe
rformance through a set of learning tasks. Since our bound holds for standard IP
S, we are able to provide insight into when regularizing IPS is useful. Namely,
we identify cases where regularization might not be needed. This goes against th
e belief that, in practice, clipped IPS often enjoys favorable performance than
standard IPS in OPL.
****

Polynomial Time and Private Learning of Unbounded Gaussian Mixture Models
Jamil Arbas, Hassan Ashtiani, Christopher Liaw
We study the problem of privately estimating the parameters of d-dimensional G
aussian Mixture Models (GMMs) with k components. For this, we develop a techni
que to reduce the problem to its non-private counterpart. This allows us to priv
atize existing non-private algorithms in a blackbox manner, while incurring only
 a small overhead in the sample complexity and running time. As the main applica
tion of our framework, we develop an (ε, δ)-differentially priv
ate algorithm to learn GMMs using the non-private algorithm of Moitra and Valian
t (2010) as a blackbox. Consequently, this gives the first sample complexity upp
er bound and first polynomial time algorithm for privately learning GMMs without
 any boundedness assumptions on the parameters. As part of our analysis, we prov
e a tight (up to a constant factor) lower bound on the total variation distance
of high-dimensional Gaussians which can be of independent interest.
****

Principled Acceleration of Iterative Numerical Methods Using Machine Learning
Sohei Arisaka, Qianxiao Li
Iterative methods are ubiquitous in large-scale scientific computing application
s, and a number of approaches based on meta-learning have been recently proposed
 to accelerate them. However, a systematic study of these approaches and how the
y differ from meta-learning is lacking. In this paper, we propose a framework to
 analyze such learning-based acceleration approaches, where one can immediately
identify a departure from classical meta-learning. We theoretically show that th
is departure may lead to arbitrary deterioration of model performance, and at th
e same time, we identify a methodology to ameliorate it by modifying the loss ob
jective, leading to a novel training method for learning-based acceleration of i
terative algorithms. We demonstrate the significant advantage and versatility of
 the proposed approach through various numerical applications.
****

Faster Rates of Convergence to Stationary Points in Differentially Private Optim
ization
Raman Arora, Raef Bassily, Tomás González, Cristóbal A Guzmán, Michael Menart, E
nayat Ullah
We study the problem of approximating stationary points of Lipschitz and smooth
functions under (ε,δ)-differential privacy (DP) in both the fin
ite-sum and stochastic settings. A point $\widehat{w}$ is called an α-sta
tionary point of a function $F:\mathbb{R}^d\rightarrow\mathbb{R}$ if $\|\nabla F
(\widehat{w})\|\leq \alpha$. We give a new construction that improves over the e
xisting rates in the stochastic optimization setting, where the goal is to find
approximate stationary points of the population risk given n samples. Our cons
truction finds a $\tilde{O}\big(\frac{1}{n^{1/3}} + \big[\frac{\sqrt{d}}{n\varep
silon}\big]^{1/2}\big)$-stationary point of the population risk in time linear i
n n. We also provide an efficient algorithm that finds an $\tilde{O}\big(\big[
\frac{\sqrt{d}}{n\varepsilon}\big]^{2/3}\big)$-stationary point in the finite-su
m setting. This improves on the previous best rate of $\tilde{O}\big(\big[\frac{
\sqrt{d}}{n\varepsilon}\big]^{1/2}\big)$. Furthermore, under the additional assu
mption of convexity, we completely characterize the sample complexity of finding
 stationary points of the population risk (up to polylog factors) and show that
the optimal rate on population stationarity is $\tilde \Theta\big(\frac{1}{\sqrt
{n}}+\frac{\sqrt{d}}{n\varepsilon}\big)$. Finally, we show that our methods can
be used to provide dimension-independent rates of $O\big(\frac{1}{\sqrt{n}}+\min
\big(\big[\frac{\sqrt{rank}}{n\varepsilon}\big]^{2/3},\frac{1}{(n\varepsilon)^{2
/5}}\big)\big)$ on population stationarity for Generalized Linear Models (GLM),

where $rank$ is the rank of the design matrix, which improves upon the previous best known rate.
****

Prototype-Sample Relation Distillation: Towards Replay-Free Continual Learning
Nader Asadi, Mohammadreza Davari, Sudhir Mudur, Rahaf Aljundi, Eugene Belilovsky
In Continual learning (CL) balancing effective adaptation while combating catastrophic forgetting is a central challenge. Many of the recent best-performing methods utilize various forms of prior task data, e.g. a replay buffer, to tackle the catastrophic forgetting problem. Having access to previous task data can be restrictive in many real-world scenarios, for example when task data is sensitive or proprietary. To overcome the necessity of using previous tasks' data, in this work, we start with strong representation learning methods that have been shown to be less prone to forgetting. We propose a holistic approach to jointly learn the representation and class prototypes while maintaining the relevance of old class prototypes and their embedded similarities. Specifically, samples are mapped to an embedding space where the representations are learned using a supervised contrastive loss. Class prototypes are evolved continually in the same latent space, enabling learning and prediction at any point. To continually adapt the prototypes without keeping any prior task data, we propose a novel distillation loss that constrains class prototypes to maintain relative similarities as compared to new task data. This method yields state-of-the-art performance in the task-incremental setting, outperforming methods relying on large amounts of data, and provides strong performance in the class-incremental setting without using any stored data points.
****

Near-Optimal Algorithms for Private Online Optimization in the Realizable Regime
Hilal Asi, Vitaly Feldman, Tomer Koren, Kunal Talwar
We consider online learning problems in the realizable setting, where there is a zero-loss solution, and propose new Differentially Private (DP) algorithms that obtain near-optimal regret bounds. For the problem of online prediction from experts, we design new algorithms that obtain near-optimal regret $O \big(\varepsilon^{-1} \mathsf{poly}(\log{d}) \big)$ where d is the number of experts. This significantly improves over the best existing regret bounds for the DP non-realizable setting which are $O \big(\varepsilon^{-1} \min\big\{d, \sqrt{T\log d}\big\} \big)$. We also develop an adaptive algorithm for the small-loss setting with regret $(L^\star+ \varepsilon^{-1}) \cdot O(\mathsf{poly}(\log{d}))$ where $L^\star$ is the total loss of the best expert. Additionally, we consider DP online convex optimization in the realizable setting and propose an algorithm with near-optimal regret $O \big(\varepsilon^{-1} \mathsf{poly}(d) \big)$, as well as an algorithm for the smooth case with regret $O \big((\sqrt{Td}/\varepsilon)^{2/3} \big)$, both significantly improving over existing bounds in the non-realizable regime.
****

From Robustness to Privacy and Back
Hilal Asi, Jonathan Ullman, Lydia Zakynthinou
We study the relationship between two desiderata of algorithms in statistical inference and machine learning—differential privacy and robustness to adversarial data corruptions. Their conceptual similarity was first observed by Dwork and Lei (STOC 2009), who observed that private algorithms satisfy robustness, and gave a general method for converting robust algorithms to private ones. However, all general methods for transforming robust algorithms into private ones lead to suboptimal error rates. Our work gives the first black-box transformation that converts any adversarially robust algorithm into one that satisfies pure differential privacy. Moreover, we show that for any low-dimensional estimation task, applying our transformation to an optimal robust estimator results in an optimal private estimator. Thus, we conclude that for any low-dimensional task, the optimal error rate for ε-differentially private estimators is essentially the same as the optimal error rate for estimators that are robust to adversarially corrupting $1/\varepsilon$ training samples. We apply our transformation to obtain new optimal private estimators for several high-dimensional statistical tas

ks, including Gaussian linear regression and PCA. Finally, we present an extensi
on of our transformation that leads to approximately differentially private algo
rithms whose error does not depend on the range of the output space, which is im
possible under pure differential privacy.
****

SGD with AdaGrad Stepsizes: Full Adaptivity with High Probability to Unknown Par
ameters, Unbounded Gradients and Affine Variance
Amit Attia, Tomer Koren
We study Stochastic Gradient Descent with AdaGrad stepsizes: a popular adaptive
(self-tuning) method for first-order stochastic optimization. Despite being well
 studied, existing analyses of this method suffer from various shortcomings: the
y either assume some knowledge of the problem parameters, impose strong global L
ipschitz conditions, or fail to give bounds that hold with high probability. We
provide a comprehensive analysis of this basic method without any of these limit
ations, in both the convex and non-convex (smooth) cases, that additionally supp
orts a general “affine variance” noise model and provides sharp rates of converg
ence in both the low-noise and high-noise regimes.
****

Adversarially Robust PAC Learnability of Real-Valued Functions
Idan Attias, Steve Hanneke
We study robustness to test-time adversarial attacks in the regression setting w
ith ℓ_p losses and arbitrary perturbation sets. We address the question of
which function classes are PAC learnable in this setting. We show that classes o
f finite fat-shattering dimension are learnable in both the realizable and agnos
tic settings. Moreover, for convex function classes, they are even properly lear
nable. In contrast, some non-convex function classes provably require improper l
earning algorithms. Our main technique is based on a construction of an adversar
ially robust sample compression scheme of a size determined by the fat-shatterin
g dimension. Along the way, we introduce a novel agnostic sample compression sch
eme for real-valued functions, which may be of independent interest.
****

Infusing Lattice Symmetry Priors in Attention Mechanisms for Sample-Efficient Ab
stract Geometric Reasoning
Mattia Atzeni, Mrinmaya Sachan, Andreas Loukas
The Abstraction and Reasoning Corpus (ARC) (Chollet, 2019) and its most recent l
anguage-complete instantiation (LARC) has been postulated as an important step t
owards general AI. Yet, even state-of-the-art machine learning models struggle t
o achieve meaningful performance on these problems, falling behind non-learning
based approaches. We argue that solving these tasks requires extreme generalizat
ion that can only be achieved by proper accounting for core knowledge priors. As
 a step towards this goal, we focus on geometry priors and introduce LatFormer,
a model that incorporates lattice symmetry priors in attention masks. We show th
at, for any transformation of the hypercubic lattice, there exists a binary atte
ntion mask that implements that group action. Hence, our study motivates a modif
ication to the standard attention mechanism, where attention weights are scaled
using soft masks generated by a convolutional network. Experiments on synthetic
geometric reasoning show that LatFormer requires 2 orders of magnitude fewer dat
a than standard attention and transformers. Moreover, our results on ARC and LAR
C tasks that incorporate geometric priors provide preliminary evidence that thes
e complex datasets do not lie out of the reach of deep learning models.
****

Learning to Initiate and Reason in Event-Driven Cascading Processes
Yuval Atzmon, Eli Meirom, Shie Mannor, Gal Chechik
Training agents to control a dynamic environment is a fundamental task in AI. In
 many environments, the dynamics can be summarized by a small set of events that
 capture the semantic behavior of the system. Typically, these events form chain
s or cascades. We often wish to change the system behavior using a single interv
ention that propagates through the cascade. For instance, one may trigger a bioc
hemical cascade to switch the state of a cell or, in logistics, reroute a truck
to meet an unexpected, urgent delivery. We introduce a new supervised learning s

etup called Cascade. An agent observes a system with known dynamics evolving fro m some initial state. The agent is given a structured semantic instruction and n eeds to make an intervention that triggers a cascade of events, such that the sy stem reaches an alternative (counterfactual) behavior. We provide a test-bed for this problem, consisting of physical objects. We combine semantic tree search w ith an event-driven forward model and devise an algorithm that learns to efficie ntly search in exponentially large semantic trees. We demonstrate that our appro ach learns to follow instructions to intervene in new complex scenes. When provi ded with an observed cascade of events, it can also reason about alternative out comes.
****

On the convergence of the MLE as an estimator of the learning rate in the Exp3 a lgorithm
Julien Aubert, Luc Lehéricy, Patricia Reynaud-Bouret
When fitting the learning data of an individual to algorithm-like learning model s, the observations are so dependent and non-stationary that one may wonder what the classical Maximum Likelihood Estimator (MLE) could do, even if it is the us ual tool applied to experimental cognition. Our objective in this work is to sho w that the estimation of the learning rate cannot be efficient if the learning r ate is constant in the classical Exp3 (Exponential weights for Exploration and E xploitation) algorithm. Secondly, we show that if the learning rate decreases po lynomially with the sample size, then the prediction error and in some cases the estimation error of the MLE satisfy bounds in probability that decrease at a po lynomial rate.
****

Dirichlet Diffusion Score Model for Biological Sequence Generation
Pavel Avdeyev, Chenlai Shi, Yuhao Tan, Kseniia Dudnyk, Jian Zhou
Designing biological sequences is an important challenge that requires satisfyin g complex constraints and thus is a natural problem to address with deep generat ive modeling. Diffusion generative models have achieved considerable success in many applications. Score-based generative stochastic differential equations (SDE) model is a continuous-time diffusion model framework that enjoys many benefits , but the originally proposed SDEs are not naturally designed for modeling discr ete data. To develop generative SDE models for discrete data such as biological sequences, here we introduce a diffusion process defined in the probability simp lex space with stationary distribution being the Dirichlet distribution. This ma kes diffusion in continuous space natural for modeling discrete data. We refer t o this approach as Dirchlet diffusion score model. We demonstrate that this tech nique can generate samples that satisfy hard constraints using a Sudoku generati on task. This generative model can also solve Sudoku, including hard puzzles, wi thout additional training. Finally, we applied this approach to develop the firs t human promoter DNA sequence design model and showed that designed sequences sh are similar properties with natural promoter sequences.
****

Gradient Descent Converges Linearly for Logistic Regression on Separable Data
Kyriakos Axiotis, Maxim Sviridenko
We show that running gradient descent with variable learning rate guarantees los s $f(x) \leq 1.1 \cdot f(x^*)+\epsilon$ for the logistic regression objective, wher e the error ϵ decays exponentially with the number of iterations and po lynomially with the magnitude of the entries of an arbitrary fixed solution x. This is in contrast to the common intuition that the absence of strong convexit y precludes linear convergence of first-order methods, and highlights the import ance of variable learning rates for gradient descent. We also apply our ideas to sparse logistic regression, where they lead to an exponential improvement of th e sparsity-error tradeoff.
****

Naive imputation implicitly regularizes high-dimensional linear models
Alexis Ayme, Claire Boyer, Aymeric Dieuleveut, Erwan Scornet
Two different approaches exist to handle missing values for prediction: either i mputation, prior to fitting any predictive algorithms, or dedicated methods able

to natively incorporate missing values. While imputation is widely (and easily) use, it is unfortunately biased when low-capacity predictors (such as linear models) are applied afterward. However, in practice, naive imputation exhibits good predictive performance. In this paper, we study the impact of imputation in a high-dimensional linear model with MCAR missing data. We prove that zero imputation performs an implicit regularization closely related to the ridge method, often used in high-dimensional problems. Leveraging on this connection, we establish that the imputation bias is controlled by a ridge bias, which vanishes in high dimension. As a predictor, we argue in favor of the averaged SGD strategy, applied to zero-imputed data. We establish an upper bound on its generalization error, highlighting that imputation is benign in the $d \gg \sqrt{n}$ regime. Experiments illustrate our findings.

****

Half-Hop: A graph upsampling approach for slowing down message passing

Mehdi Azabou, Venkataramana Ganesh, Shantanu Thakoor, Chi-Heng Lin, Lakshmi Sathidevi, Ran Liu, Michal Valko, Petar Veli■kovi■, Eva L Dyer

Message passing neural networks have shown a lot of success on graph-structured data. However, there are many instances where message passing can lead to over-smoothing or fail when neighboring nodes belong to different classes. In this work, we introduce a simple yet general framework for improving learning in message passing neural networks. Our approach essentially upsamples edges in the original graph by adding "slow nodes" at each edge that can mediate communication between a source and a target node. Our method only modifies the input graph, making it plug-and-play and easy to use with existing models. To understand the benefits of slowing down message passing, we provide theoretical and empirical analyses. We report results on several supervised and self-supervised benchmarks, and show improvements across the board, notably in heterophilic conditions where adjacent nodes are more likely to have different labels. Finally, we show how our approach can be used to generate augmentations for self-supervised learning, where slow nodes are randomly introduced into different edges in the graph to generate multi-scale views with variable path lengths.

****

CLUTR: Curriculum Learning via Unsupervised Task Representation Learning

Abdus Salam Azad, Izzeddin Gur, Jasper Emhoff, Nathaniel Alexis, Aleksandra Faust, Pieter Abbeel, Ion Stoica

Reinforcement Learning (RL) algorithms are often known for sample inefficiency and difficult generalization. Recently, Unsupervised Environment Design (UED) emerged as a new paradigm for zero-shot generalization by simultaneously learning a task distribution and agent policies on the generated tasks. This is a non-stationary process where the task distribution evolves along with agent policies; creating an instability over time. While past works demonstrated the potential of such approaches, sampling effectively from the task space remains an open challenge, bottlenecking these approaches. To this end, we introduce CLUTR: a novel unsupervised curriculum learning algorithm that decouples task representation and curriculum learning into a two-stage optimization. It first trains a recurrent variational autoencoder on randomly generated tasks to learn a latent task manifold. Next, a teacher agent creates a curriculum by maximizing a minimax REGRET-based objective on a set of latent tasks sampled from this manifold. Using the fixed-pretrained task manifold, we show that CLUTR successfully overcomes the non-stationarity problem and improves stability. Our experimental results show CLUTR outperforms PAIRED, a principled and popular UED method, in the challenging CarRacing and navigation environments: achieving 10.6X and 45% improvement in zero-shot generalization, respectively. CLUTR also performs comparably to the non-UED state-of-the-art for CarRacing, while requiring 500X fewer environment interactions. We open source our code at https://github.com/clutr/clutr.

****

Personalized Subgraph Federated Learning

Jinheon Baek, Wonyong Jeong, Jiongdao Jin, Jaehong Yoon, Sung Ju Hwang

Subgraphs of a larger global graph may be distributed across multiple devices, and only locally accessible due to privacy restrictions, although there may be li

nks between subgraphs. Recently proposed subgraph Federated Learning (FL) methods deal with those missing links across local subgraphs while distributively training Graph Neural Networks (GNNs) on them. However, they have overlooked the inevitable heterogeneity between subgraphs comprising different communities of a global graph, consequently collapsing the incompatible knowledge from local GNN models. To this end, we introduce a new subgraph FL problem, personalized subgraph FL, which focuses on the joint improvement of the interrelated local GNNs rather than learning a single global model, and propose a novel framework, FEDerated Personalized sUBgraph learning (FED-PUB), to tackle it. Since the server cannot access the subgraph in each client, FED-PUB utilizes functional embeddings of the local GNNs using random graphs as inputs to compute similarities between them, and use the similarities to perform weighted averaging for server-side aggregation. Further, it learns a personalized sparse mask at each client to select and update only the subgraph-relevant subset of the aggregated parameters. We validate our FED-PUB for its subgraph FL performance on six datasets, considering both non-overlapping and overlapping subgraphs, on which it significantly outperforms relevant baselines. Our code is available at https://github.com/JinheonBaek/FED-PUB.

****

Efficient Self-supervised Learning with Contextualized Target Representations for Vision, Speech and Language

Alexei Baevski, Arun Babu, Wei-Ning Hsu, Michael Auli

Current self-supervised learning algorithms are often modality-specific and require large amounts of computational resources. To address these issues, we increase the training efficiency of data2vec, a learning objective that generalizes across several modalities. We do not encode masked tokens, use a fast convolutional decoder and amortize the effort to build teacher representations. data2vec 2.0 benefits from the rich contextualized target representations introduced in data2vec which enable a fast self-supervised learner. Experiments on ImageNet-1K image classification show that data2vec 2.0 matches the accuracy of Masked Autoencoders in 16.4x lower pre-training time, on Librispeech speech recognition it performs as well as wav2vec 2.0 in 10.6x less time, and on GLUE natural language understanding it matches a retrained RoBERTa model in half the time. Trading some speed for accuracy results in ImageNet-1K top-1 accuracy of 86.8% with a ViT-L model trained for 150 epochs.

****

Efficient preconditioned stochastic gradient descent for estimation in latent variable models

Charlotte Baey, Maud Delattre, Estelle Kuhn, Jean-Benoist Leger, Sarah Lemler

Latent variable models are powerful tools for modeling complex phenomena involving in particular partially observed data, unobserved variables or underlying complex unknown structures. Inference is often difficult due to the latent structure of the model. To deal with parameter estimation in the presence of latent variables, well-known efficient methods exist, such as gradient-based and EM-type algorithms, but with practical and theoretical limitations. In this paper, we propose as an alternative for parameter estimation an efficient preconditioned stochastic gradient algorithm. Our method includes a preconditioning step based on a positive definite Fisher information matrix estimate. We prove convergence results for the proposed algorithm under mild assumptions for very general latent variables models. We illustrate through relevant simulations the performance of the proposed methodology in a nonlinear mixed effects model and in a stochastic block model.

****

Feed Two Birds with One Scone: Exploiting Wild Data for Both Out-of-Distribution Generalization and Detection

Haoyue Bai, Gregory Canal, Xuefeng Du, Jeongyeol Kwon, Robert D Nowak, Yixuan Li

Modern machine learning models deployed in the wild can encounter both covariate and semantic shifts, giving rise to the problems of out-of-distribution (OOD) generalization and OOD detection respectively. While both problems have received significant research attention lately, they have been pursued independently. Thi

s may not be surprising, since the two tasks have seemingly conflicting goals. This paper provides a new unified approach that is capable of simultaneously generalizing to covariate shifts while robustly detecting semantic shifts. We propose a margin-based learning framework that exploits freely available unlabeled data in the wild that captures the environmental test-time OOD distributions under both covariate and semantic shifts. We show both empirically and theoretically that the proposed margin constraint is the key to achieving both OOD generalization and detection. Extensive experiments show the superiority of our framework, outperforming competitive baselines that specialize in either OOD generalization or OOD detection. Code is publicly available at https://github.com/deeplearning-wisc/scone.

****

Answering Complex Logical Queries on Knowledge Graphs via Query Computation Tree Optimization

Yushi Bai, Xin Lv, Juanzi Li, Lei Hou

Answering complex logical queries on incomplete knowledge graphs is a challenging task, and has been widely studied. Embedding-based methods require training on complex queries and may not generalize well to out-of-distribution query structures. Recent work frames this task as an end-to-end optimization problem, and it only requires a pretrained link predictor. However, due to the exponentially large combinatorial search space, the optimal solution can only be approximated, limiting the final accuracy. In this work, we propose QTO (Query Computation Tree Optimization) that can efficiently find the exact optimal solution. QTO finds the optimal solution by a forward-backward propagation on the tree-like computation graph, i.e., query computation tree. In particular, QTO utilizes the independence encoded in the query computation tree to reduce the search space, where only local computations are involved during the optimization procedure. Experiments on 3 datasets show that QTO obtains state-of-the-art performance on complex query answering, outperforming previous best results by an average of 22%. Moreover, QTO can interpret the intermediate solutions for each of the one-hop atoms in the query with over 90% accuracy.

****

Linear optimal partial transport embedding

Yikun Bai, Ivan Vladimir Medri, Rocio Diaz Martin, Rana Shahroz, Soheil Kolouri

Optimal transport (OT) has gained popularity due to its various applications in fields such as machine learning, statistics, and signal processing. However, the balanced mass requirement limits its performance in practical problems. To address these limitations, variants of the OT problem, including unbalanced OT, Optimal partial transport (OPT), and Hellinger Kantorovich (HK), have been proposed. In this paper, we propose the Linear optimal partial transport (LOPT) embedding, which extends the (local) linearization technique on OT and HK to the OPT problem. The proposed embedding allows for faster computation of OPT distance between pairs of positive measures. Besides our theoretical contributions, we demonstrate the LOPT embedding technique in point-cloud interpolation and PCA analysis. Our code is available at https://github.com/Baio0/LinearOPT.

****

Implicit Graph Neural Networks: A Monotone Operator Viewpoint

Justin Baker, Qingsong Wang, Cory D Hauck, Bao Wang

Implicit graph neural networks (IGNNs) – that solve a fixed-point equilibrium equation using Picard iteration for representation learning – have shown remarkable performance in learning long-range dependencies (LRD) in the underlying graphs. However, IGNNs suffer from several issues, including 1) their expressivity is limited by their parameterizations for the well-posedness guarantee, 2) IGNNs are unstable in learning LRD, and 3) IGNNs become computationally inefficient when learning LRD. In this paper, we provide a new well-posedness characterization for IGNNs leveraging monotone operator theory, resulting in a much more expressive parameterization than the existing one. We also propose an orthogonal parameterization for IGNN based on Cayley transform to stabilize learning LRD. Furthermore, we leverage Anderson-accelerated operator splitting schemes to efficiently solve for the fixed point of the equilibrium equation of IGNN with monotone or or

thogonal parameterization. We verify the computational efficiency and accuracy o
f the new models over existing IGNNs on various graph learning tasks at both gra
ph and node levels.
****

Tensor Decompositions Meet Control Theory: Learning General Mixtures of Linear D
ynamical Systems
Ainesh Bakshi, Allen Liu, Ankur Moitra, Morris Yau
Recently Chen and Poor initiated the study of learning mixtures of linear dynami
cal systems. While linear dynamical systems already have wide-ranging applicatio
ns in modeling time-series data, using mixture models can lead to a better fit o
r even a richer understanding of underlying subpopulations represented in the da
ta. In this work we give a new approach to learning mixtures of linear dynamical
 systems that is based on tensor decompositions. As a result, our algorithm succ
eeds without strong separation conditions on the components, and can be used to
compete with the Bayes optimal clustering of the trajectories. Moreover our algo
rithm works in the challenging partially-observed setting. Our starting point is
 the simple but powerful observation that the classic Ho-Kalman algorithm is a r
elative of modern tensor decomposition methods for learning latent variable mode
ls. This gives us a playbook for how to extend it to work with more complicated
generative models.
****

Block Subsampled Randomized Hadamard Transform for Nyström Approximation on Dist
ributed Architectures
Oleg Balabanov, Matthias Beaupère, Laura Grigori, Victor Lederer
This article introduces a novel structured random matrix composed blockwise from
 subsampled randomized Hadamard transforms (SRHTs). The block SRHT is expected t
o outperform well-known dimension reduction maps, including SRHT and Gaussian ma
trices on distributed architectures. We prove that a block SRHT with enough rows
 is an oblivious subspace embedding, i.e., an approximate isometry for an arbitr
ary low-dimensional subspace with high probability. Our estimate of the required
 number of rows is similar to that of the standard SRHT. This suggests that the
two transforms should provide the same accuracy of approximation in the algorith
ms. The block SRHT can be readily incorporated into randomized methods for compu
ting a low-rank approximation of a large-scale matrix, such as the Nyström metho
d. For completeness, we revisit this method with a discussion of its implementat
ion on distributed architectures.
****

Efficient Online Reinforcement Learning with Offline Data
Philip J. Ball, Laura Smith, Ilya Kostrikov, Sergey Levine
Sample efficiency and exploration remain major challenges in online reinforcemen
t learning (RL). A powerful approach that can be applied to address these issues
 is the inclusion of offline data, such as prior trajectories from a human exper
t or a sub-optimal exploration policy. Previous methods have relied on extensive
 modifications and additional complexity to ensure the effective use of this dat
a. Instead, we ask: can we simply apply existing off-policy methods to leverage
offline data when learning online? In this work, we demonstrate that the answer
is yes; however, a set of minimal but important changes to existing off-policy R
L algorithms are required to achieve reliable performance. We extensively ablate
 these design choices, demonstrating the key factors that most affect performanc
e, and arrive at a set of recommendations that practitioners can readily apply,
whether their data comprise a small number of expert demonstrations or large vol
umes of sub-optimal trajectories. We see that correct application of these simpl
e recommendations can provide a $\mathbf{2.5\times}$ improvement over existing a
pproaches across a diverse set of competitive benchmarks, with no additional com
putational overhead.
****

Mirror Sinkhorn: Fast Online Optimization on Transport Polytopes
Marin Ballu, Quentin Berthet
Optimal transport is an important tool in machine learning, allowing to capture
geometric properties of the data through a linear program on transport polytopes

. We present a single-loop optimization algorithm for minimizing general convex objectives on these domains, utilizing the principles of Sinkhorn matrix scaling and mirror descent. The proposed algorithm is robust to noise, and can be used in an online setting. We provide theoretical guarantees for convex objectives and experimental results showcasing it effectiveness on both synthetic and real-world data.

****

On the Functional Similarity of Robust and Non-Robust Neural Representations

András Balogh, Márk Jelasity

Model stitching—where the internal representations of two neural networks are aligned linearly—helped demonstrate that the representations of different neural networks for the same task are surprisingly similar in a functional sense. At the same time, the representations of adversarially robust networks are considered to be different from non-robust representations. For example, robust image classifiers are invertible, while non-robust networks are not. Here, we investigate the functional similarity of robust and non-robust representations for image classification with the help of model stitching. We find that robust and non-robust networks indeed have different representations. However, these representations are compatible regarding accuracy. From the point of view of robust accuracy, compatibility decreases quickly after the first few layers but the representations become compatible again in the last layers, in the sense that the properties of the front model can be recovered. Moreover, this is true even in the case of cross-task stitching. Our results suggest that stitching in the initial, preprocessing layers and the final, abstract layers test different kinds of compatibilities. In particular, the final layers are easy to match, because their representations depend mostly on the same abstract task specification, in our case, the classification of the input into n classes.

****

Robust Budget Pacing with a Single Sample

Santiago R. Balseiro, Rachitesh Kumar, Vahab Mirrokni, Balasubramanian Sivan, Di Wang

Major Internet advertising platforms offer budget pacing tools as a standard service for advertisers to manage their ad campaigns. Given the inherent non-stationarity in an advertiser’s value and also competing advertisers’ values over time, a commonly used approach is to learn a target expenditure plan that specifies a target spend as a function of time, and then run a controller that tracks this plan. This raises the question: how many historical samples are required to learn a good expenditure plan? We study this question by considering an advertiser repeatedly participating in T second-price auctions, where the tuple of her value and the highest competing bid is drawn from an unknown time-varying distribution. The advertiser seeks to maximize her total utility subject to her budget constraint. Prior work has shown the sufficiency of $T\log T$ samples per distribution to achieve the optimal $O(\sqrt{T})$-regret. We dramatically improve this state-of-the-art and show that just one sample per distribution is enough to achieve the near-optimal $\tilde O(\sqrt{T})$-regret, while still being robust to noise in the sampling distributions.

****

Dynamic Constrained Submodular Optimization with Polylogarithmic Update Time

Kiarash Banihashem, Leyla Biabani, Samira Goudarzi, Mohammadtaghi Hajiaghayi, Peyman Jabbarzade, Morteza Monemizadeh

Maximizing a monotone submodular function under cardinality constraint k is a core problem in machine learning and database with many basic applications, including video and data summarization, recommendation systems, feature extraction, exemplar clustering, and coverage problems. We study this classic problem in the fully dynamic model where a stream of insertions and deletions of elements of an underlying ground set is given and the goal is to maintain an approximate solution using a fast update time. A recent paper at NeurIPS’20 by Lattanzi, Mitrovic, Norouzi-Fard, Tarnawski, Zadimoghaddam claims to obtain a dynamic algorithm for this problem with a $(\frac{1}{2} -\epsilon)$ approximation ratio and a query complexity bounded by $\mathrm{poly}(\log(n),\log(k),\epsilon^{-1})$. However,

as we explain in this paper, the analysis has some important gaps. Having a dyna mic algorithm for the problem with polylogarithmic update time is even more impo rtant in light of a recent result by Chen and Peng at STOC’22 who show a matchin g lower bound for the problem – any randomized algorithm with a $\frac{1}{2}+\ep silon$ approximation ratio must have an amortized query complexity that is polyn omial in n. In this paper, we develop a simpler algorithm for the problem that maintains a $(\frac{1}{2}-\epsilon)$-approximate solution for submodular maximi zation under cardinality constraint k using a polylogarithmic amortized update time.
****
One Transformer Fits All Distributions in Multi-Modal Diffusion at Scale
Fan Bao, Shen Nie, Kaiwen Xue, Chongxuan Li, Shi Pu, Yaole Wang, Gang Yue, Yue C ao, Hang Su, Jun Zhu
This paper proposes a unified diffusion framework (dubbed UniDiffuser) to fit al l distributions relevant to a set of multi-modal data in one model. Our key insi ght is – learning diffusion models for marginal, conditional, and joint distribu tions can be unified as predicting the noise in the perturbed data, where the pe rturbation levels (i.e. timesteps) can be different for different modalities. In spired by the unified view, UniDiffuser learns all distributions simultaneously with a minimal modification to the original diffusion model – perturbs data in a ll modalities instead of a single modality, inputs individual timesteps in diffe rent modalities, and predicts the noise of all modalities instead of a single mo dality. UniDiffuser is parameterized by a transformer for diffusion models to ha ndle input types of different modalities. Implemented on large-scale paired imag e-text data, UniDiffuser is able to perform image, text, text-to-image, image-to -text, and image-text pair generation by setting proper timesteps without additi onal overhead. In particular, UniDiffuser is able to produce perceptually realis tic samples in all tasks and its quantitative results (e.g., the FID and CLIP sc ore) are not only superior to existing general-purpose models but also comparabl e to the bespoken models (e.g., Stable Diffusion and DALL-E 2) in representative tasks (e.g., text-to-image generation).
****
Optimizing the Collaboration Structure in Cross-Silo Federated Learning
Wenxuan Bao, Haohan Wang, Jun Wu, Jingrui He
In federated learning (FL), multiple clients collaborate to train machine learni ng models together while keeping their data decentralized. Through utilizing mor e training data, FL suffers from the potential negative transfer problem: the gl obal FL model may even perform worse than the models trained with local data onl y. In this paper, we propose FedCollab, a novel FL framework that alleviates neg ative transfer by clustering clients into non-overlapping coalitions based on th eir distribution distances and data quantities. As a result, each client only co llaborates with the clients having similar data distributions, and tends to coll aborate with more clients when it has less data. We evaluate our framework with a variety of datasets, models, and types of non-IIDness. Our results demonstrate that FedCollab effectively mitigates negative transfer across a wide range of F L algorithms and consistently outperforms other clustered FL algorithms.
****
MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation
Omer Bar-Tal, Lior Yariv, Yaron Lipman, Tali Dekel
Recent advances in text-to-image generation with diffusion models present transf ormative capabilities in image quality. However, user controllability of the gen erated image, and fast adaptation to new tasks still remains an open challenge, currently mostly addressed by costly and long re-training and fine-tuning or ad-hoc adaptations to specific image generation tasks. In this work, we present Mul tiDiffusion, a unified framework that enables versatile and controllable image g eneration, using a pre-trained text-to-image diffusion model, without any furthe r training or finetuning. At the center of our approach is a new generation proc ess, based on an optimization task that binds together multiple diffusion genera tion processes with a shared set of parameters or constraints. We show that Mult iDiffusion can be readily applied to generate high quality and diverse images th

at adhere to user-provided controls, such as desired aspect ratio (e.g., panoram
a), and spatial guiding signals, ranging from tight segmentation masks to boundi
ng boxes.
****

Reinforcement Learning with General Utilities: Simpler Variance Reduction and La
rge State-Action Space

Anas Barakat, Ilyas Fatkhullin, Niao He

We consider the reinforcement learning (RL) problem with general utilities which
 consists in maximizing a function of the state-action occupancy measure. Beyond
 the standard cumulative reward RL setting, this problem includes as particular
cases constrained RL, pure exploration and learning from demonstrations among ot
hers. For this problem, we propose a simpler single-loop parameter-free normaliz
ed policy gradient algorithm. Implementing a recursive momentum variance reducti
on mechanism, our algorithm achieves $\tilde{\mathcal{O}}(\epsilon^{-3})$ and $\
tilde{\mathcal{O}}(\epsilon^{-2})$ sample complexities for ϵ-first-orde
r stationarity and ϵ-global optimality respectively, under adequate ass
umptions. We further address the setting of large finite state action spaces via
 linear function approximation of the occupancy measure and show a $\tilde{\math
cal{O}}(\epsilon^{-4})$ sample complexity for a simple policy gradient method wi
th a linear regression subroutine.
****

Interpretable Neural-Symbolic Concept Reasoning

Pietro Barbiero, Gabriele Ciravegna, Francesco Giannini, Mateo Espinosa Zarlenga
, Lucie Charlotte Magister, Alberto Tonda, Pietro Lio, Frederic Precioso, Mateja
 Jamnik, Giuseppe Marra

Deep learning methods are highly accurate, yet their opaque decision process pre
vents them from earning full human trust. Concept-based models aim to address th
is issue by learning tasks based on a set of human-understandable concepts. Howe
ver, state-of-the-art concept-based models rely on high-dimensional concept embe
dding representations which lack a clear semantic meaning, thus questioning the
interpretability of their decision process. To overcome this limitation, we prop
ose the Deep Concept Reasoner (DCR), the first interpretable concept-based model
 that builds upon concept embeddings. In DCR, neural networks do not make task p
redictions directly, but they build syntactic rule structures using concept embe
ddings. DCR then executes these rules on meaningful concept truth degrees to pro
vide a final interpretable and semantically-consistent prediction in a different
iable manner. Our experiments show that DCR: (i) improves up to +25% w.r.t. stat
e-of-the-art interpretable concept-based models on challenging benchmarks (ii) d
iscovers meaningful logic rules matching known ground truths even in the absence
 of concept supervision during training, and (iii), facilitates the generation o
f counterfactual examples providing the learnt rules as guidance.
****

Moccasin: Efficient Tensor Rematerialization for Neural Networks

Burak Bartan, Haoming Li, Harris Teague, Christopher Lott, Bistra Dilkina

The deployment and training of neural networks on edge computing devices pose ma
ny challenges. The low memory nature of edge devices is often one of the biggest
 limiting factors encountered in the deployment of large neural network models.
Tensor rematerialization or recompute is a way to address high memory requiremen
ts for neural network training and inference. In this paper we consider the prob
lem of execution time minimization of compute graphs subject to a memory budget.
 In particular, we develop a new constraint programming formulation called Mocca
sin with only $O(n)$ integer variables, where n is the number of nodes in the
compute graph. This is a significant improvement over the works in the recent li
terature that propose formulations with $O(n^2)$ Boolean variables. We present n
umerical studies that show that our approach is up to an order of magnitude fast
er than recent work especially for large-scale graphs.
****

User-level Private Stochastic Convex Optimization with Optimal Rates

Raef Bassily, Ziteng Sun

We study the problem of differentially private (DP) stochastic convex optimizati

on (SCO) under the notion of user-level differential privacy. In this problem, t here are n users, each contributing $m>1$ samples to the input dataset of the private SCO algorithm, and the notion of indistinguishability embedded in DP is w.r.t. replacing the entire local dataset of any given user. Under smoothness co nditions of the loss, we establish the optimal rates for user-level DP-SCO in bo th the central and local models of DP. In particular, we show, roughly, that the optimal rate is $\frac{1}{\sqrt{nm}}+\frac{\sqrt{d}}{\varepsilon n \sqrt{m}}$ i n the central setting and is $\frac{\sqrt{d}}{\varepsilon \sqrt{nm}}$ in the loc al setting, where d is the dimensionality of the problem and ε is the privacy parameter. Our algorithms combine new user-level DP mean estimation techniques with carefully designed first-order stochastic optimization methods. For the central DP setting, our optimal rate improves over the rate attained for the same setting in Levy et al. (2021) by $\sqrt{d}$ factor. One of the main in gredients that enabled such an improvement is a novel application of the general ization properties of DP in the context of multi-pass stochastic gradient method s.

****

A Statistical Perspective on Retrieval-Based Models

Soumya Basu, Ankit Singh Rawat, Manzil Zaheer

Many modern high-performing machine learning models increasingly rely on scaling up models, e.g., transformer networks. Simultaneously, a parallel line of work aims to improve the model performance by augmenting an input instance with other (labeled) instances during inference. Examples of such augmentations include ta sk-specific prompts and similar examples retrieved from the training data by a n onparametric component. Despite a growing literature showcasing the promise of t hese retrieval-based models, their theoretical underpinnings %for such models re main under-explored. In this paper, we present a formal treatment of retrieval-b ased models to characterize their performance via a novel statistical perspectiv e. In particular, we study two broad classes of retrieval-based classification a pproaches: First, we analyze a local learning framework that employs an explicit local empirical risk minimization based on retrieved examples for each input in stance. Interestingly, we show that breaking down the underlying learning task i nto local sub-tasks enables the model to employ a low complexity parametric comp onent to ensure good overall performance. The second class of retrieval-based ap proaches we explore learns a global model using kernel methods to directly map a n input instance and retrieved examples to a prediction, without explicitly solv ing a local learning task.

****

Human-Timescale Adaptation in an Open-Ended Task Space

Jakob Bauer, Kate Baumli, Feryal Behbahani, Avishkar Bhoopchand, Nathalie Bradle y-Schmieg, Michael Chang, Natalie Clay, Adrian Collister, Vibhavari Dasagi, Lucy Gonzalez, Karol Gregor, Edward Hughes, Sheleem Kashem, Maria Loks-Thompson, Han nah Openshaw, Jack Parker-Holder, Shreya Pathak, Nicolas Perez-Nieves, Nemanja R akicevic, Tim Rocktäschel, Yannick Schroecker, Satinder Singh, Jakub Sygnowski, Karl Tuyls, Sarah York, Alexander Zacherl, Lei M Zhang

Foundation models have shown impressive adaptation and scalability in supervised and self-supervised learning problems, but so far these successes have not full y translated to reinforcement learning (RL). In this work, we demonstrate that t raining an RL agent at scale leads to a general in-context learning algorithm th at can adapt to open-ended novel embodied 3D problems as quickly as humans. In a vast space of held-out environment dynamics, our adaptive agent (AdA) displays on-the-fly hypothesis-driven exploration, efficient exploitation of acquired kno wledge, and can successfully be prompted with first-person demonstrations. Adapt ation emerges from three ingredients: (1) meta-reinforcement learning across a v ast, smooth and diverse task distribution, (2) a policy parameterised as a large -scale attention-based memory architecture, and (3) an effective automated curri culum that prioritises tasks at the frontier of an agent’s capabilities. We demo nstrate characteristic scaling laws with respect to network size, memory length, and richness of the training task distribution. We believe our results lay the foundation for increasingly general and adaptive RL agents that perform well acr

oss ever-larger open-ended domains.
****
A Kernel Stein Test of Goodness of Fit for Sequential Models
Jerome Baum, Heishiro Kanagawa, Arthur Gretton
We propose a goodness-of-fit measure for probability densities modeling observations with varying dimensionality, such as text documents of differing lengths or variable-length sequences. The proposed measure is an instance of the kernel Stein discrepancy (KSD), which has been used to construct goodness-of-fit tests for unnormalized densities. The KSD is defined by its Stein operator: current operators used in testing apply to fixed-dimensional spaces. As our main contribution, we extend the KSD to the variable-dimension setting by identifying appropriate Stein operators, and propose a novel KSD goodness-of-fit test. As with the previous variants, the proposed KSD does not require the density to be normalized, allowing the evaluation of a large class of models. Our test is shown to perform well in practice on discrete sequential data benchmarks.
****
Individually Fair Learning with One-Sided Feedback
Yahav Bechavod, Aaron Roth
We consider an online learning problem with one-sided feedback, in which the learner is able to observe the true label only for positively predicted instances. On each round, k instances arrive and receive classification outcomes according to a randomized policy deployed by the learner, whose goal is to maximize accuracy while deploying individually fair policies. We first present a novel auditing scheme, capable of utilizing feedback from dynamically-selected panels of multiple, possibly inconsistent, auditors regarding fairness violations. In particular, we show how our proposed auditing scheme allows for algorithmically exploring the resulting accuracy-fairness frontier, with no need for additional feedback from auditors. We then present an efficient reduction from our problem of online learning with one-sided feedback and a panel reporting fairness violations to the contextual combinatorial semi-bandit problem (Cesa-Bianchi & Lugosi, 2009; Gyorgy et al., 2007), allowing us to leverage algorithms for contextual combinatorial semi-bandits to establish multi-criteria no regret guarantees in our setting, simultaneously for accuracy and fairness. Our results eliminate two potential sources of bias from prior work: the “hidden outcomes” that are not available to an algorithm operating in the full information setting, and human biases that might be present in any single human auditor, but can be mitigated by selecting a well-chosen panel.
****
Predicting Ordinary Differential Equations with Transformers
Sören Becker, Michal Klein, Alexander Neitz, Giambattista Parascandolo, Niki Kilbertus
We develop a transformer-based sequence-to-sequence model that recovers scalar ordinary differential equations (ODEs) in symbolic form from irregularly sampled and noisy observations of a single solution trajectory. We demonstrate in extensive empirical evaluations that our model performs better or on par with existing methods in terms of accurate recovery across various settings. Moreover, our method is efficiently scalable: after one-time pretraining on a large set of ODEs, we can infer the governing law of a new observed solution in a few forward passes of the model.
****
Explaining Reinforcement Learning with Shapley Values
Daniel Beechey, Thomas M. S. Smith, Özgür ■im■ek
For reinforcement learning systems to be widely adopted, their users must understand and trust them. We present a theoretical analysis of explaining reinforcement learning using Shapley values, following a principled approach from game theory for identifying the contribution of individual players to the outcome of a cooperative game. We call this general framework Shapley Values for Explaining Reinforcement Learning (SVERL). Our analysis exposes the limitations of earlier uses of Shapley values in reinforcement learning. We then develop an approach that uses Shapley values to explain agent performance. In a variety of domains, SVERL

produces meaningful explanations that match and supplement human intuition.
****
TIDE: Time Derivative Diffusion for Deep Learning on Graphs
Maysam Behmanesh, Maximilian Krahn, Maks Ovsjanikov
A prominent paradigm for graph neural networks is based on the message-passing framework. In this framework, information communication is realized only between neighboring nodes. The challenge of approaches that use this paradigm is to ensure efficient and accurate long-distance communication between nodes, as deep convolutional networks are prone to over smoothing. In this paper, we present a novel method based on time derivative graph diffusion (TIDE) to overcome these structural limitations of the message-passing framework. Our approach allows for optimizing the spatial extent of diffusion across various tasks and network channels, thus enabling medium and long-distance communication efficiently. Furthermore, we show that our architecture design also enables local message-passing and thus inherits from the capabilities of local message-passing approaches. We show that on both widely used graph benchmarks and synthetic mesh and graph datasets, the proposed framework outperforms state-of-the-art methods by a significant margin.
****
Fast as CHITA: Neural Network Pruning with Combinatorial Optimization
Riade Benbaki, Wenyu Chen, Xiang Meng, Hussein Hazimeh, Natalia Ponomareva, Zhe Zhao, Rahul Mazumder
The sheer size of modern neural networks makes model serving a serious computational challenge. A popular class of compression techniques overcomes this challenge by pruning or sparsifying the weights of pretrained networks. While useful, these techniques often face serious tradeoffs between computational requirements and compression quality. In this work, we propose a novel optimization-based pruning framework that considers the combined effect of pruning (and updating) multiple weights subject to a sparsity constraint. Our approach, CHITA, extends the classical Optimal Brain Surgeon framework and results in significant improvements in speed, memory, and performance over existing optimization-based approaches for network pruning. CHITA's main workhorse performs combinatorial optimization updates on a memory-friendly representation of local quadratic approximation(s) of the loss function. On a standard benchmark of pretrained models and datasets, CHITA leads to superior sparsity-accuracy tradeoffs than competing methods. For example, for MLPNet with only 2% of the weights retained, our approach improves the accuracy by 63% relative to the state of the art. Furthermore, when used in conjunction with fine-tuning SGD steps, our method achieves significant accuracy gains over state-of-the-art approaches. Our code is publicly available at: https://github.com/mazumder-lab/CHITA .
****
Continuously Parameterized Mixture Models
Christopher M Bender, Yifeng Shi, Marc Niethammer, Junier Oliva
Mixture models are universal approximators of smooth densities but are difficult to utilize in complicated datasets due to restrictions on typically available modes and challenges with initialiations. We show that by continuously parameterizing a mixture of factor analyzers using a learned ordinary differential equation, we can improve the fit of mixture models over direct methods. Once trained, the mixture components can be extracted and the neural ODE can be discarded, leaving us with an effective, but low-resource model. We additionally explore the use of a training curriculum from an easy-to-model latent space extracted from a normalizing flow to the more complex input space and show that the smooth curriculum helps to stabilize and improve results with and without the continuous parameterization. Finally, we introduce a hierarchical version of the model to enable more flexible, robust classification and clustering, and show substantial improvements against traditional parameterizations of GMMs.
****
Controllable Neural Symbolic Regression
Tommaso Bendinelli, Luca Biggio, Pierre-Alexandre Kamienny
In symbolic regression, the objective is to find an analytical expression that a

ccurately fits experimental data with the minimal use of mathematical symbols su ch as operators, variables, and constants. However, the combinatorial space of p ossible expressions can make it challenging for traditional evolutionary algorit hms to find the correct expression in a reasonable amount of time. To address th is issue, Neural Symbolic Regression (NSR) algorithms have been developed that c an quickly identify patterns in the data and generate analytical expressions. Ho wever, these methods, in their current form, lack the capability to incorporate user-defined prior knowledge, which is often required in natural sciences and en gineering fields. To overcome this limitation, we propose a novel neural symboli c regression method, named Neural Symbolic Regression with Hypothesis (NSRwH) th at enables the explicit incorporation of assumptions about the expected structur e of the ground-truth expression into the prediction process. Our experiments de monstrate that the proposed conditioned deep learning model outperforms its unco nditioned counterparts in terms of accuracy while also providing control over th e predicted expression structure.

****

On Second-Order Scoring Rules for Epistemic Uncertainty Quantification

Viktor Bengs, Eyke Hüllermeier, Willem Waegeman

It is well known that accurate probabilistic predictors can be trained through e mpirical risk minimisation with proper scoring rules as loss functions. While su ch learners capture so-called aleatoric uncertainty of predictions, various mach ine learning methods have recently been developed with the goal to let the learn er also represent its epistemic uncertainty, i.e., the uncertainty caused by a l ack of knowledge and data. An emerging branch of the literature proposes the use of a second-order learner that provides predictions in terms of distributions o n probability distributions. However, recent work has revealed serious theoretic al shortcomings for second-order predictors based on loss minimisation. In this paper, we generalise these findings and prove a more fundamental result: There s eems to be no loss function that provides an incentive for a second-order learne r to faithfully represent its epistemic uncertainty in the same manner as proper scoring rules do for standard (first-order) learners. As a main mathematical to ol to prove this result, we introduce the generalised notion of second-order sco ring rules.

****

Certified Robust Neural Networks: Generalization and Corruption Resistance

Amine Bennouna, Ryan Lucas, Bart Van Parys

Recent work have demonstrated that robustness (to "corruption") can be at odds w ith generalization. Adversarial training, for instance, aims to reduce the probl ematic susceptibility of modern neural networks to small data perturbations. Sur prisingly, overfitting is a major concern in adversarial training despite being mostly absent in standard training. We provide here theoretical evidence for thi s peculiar “robust overfitting” phenomenon. Subsequently, we advance a novel dis tributionally robust loss function bridging robustness and generalization. We de monstrate both theoretically as well as empirically the loss to enjoy a certifie d level of robustness against two common types of corruption|data evasion and po isoning attacks|while ensuring guaranteed generalization. We show through carefu l numerical experiments that our resulting holistic robust (HR) training procedu re yields SOTA performance. Finally, we indicate that HR training can be interpr eted as a direct extension of adversarial training and comes with a negligible a dditional computational burden. A ready-to-use python library implementing our a lgorithm is available at https://github.com/RyanLucas3/HR_Neural_Networks.

****

Gaussian processes at the Helm(holtz): A more fluid model for ocean currents

Renato Berlinghieri, Brian L. Trippe, David R. Burt, Ryan James Giordano, Kaushi k Srinivasan, Tamay Özgökmen, Junfei Xia, Tamara Broderick

Oceanographers are interested in predicting ocean currents and identifying diver gences in a current vector field based on sparse observations of buoy velocities . Since we expect current dynamics to be smooth but highly non-linear, Gaussian processes (GPs) offer an attractive model. But we show that applying a GP with a standard stationary kernel directly to buoy data can struggle at both current p

rediction and divergence identification – due to some physically unrealistic pri
or assumptions. To better reflect known physical properties of currents, we prop
ose to instead put a standard stationary kernel on the divergence and curl-free
components of a vector field obtained through a Helmholtz decomposition. We show
 that, because this decomposition relates to the original vector field just via
mixed partial derivatives, we can still perform inference given the original dat
a with only a small constant multiple of additional computational expense. We il
lustrate the benefits of our method on synthetic and real oceans data.
****

Optimal Rates and Efficient Algorithms for Online Bayesian Persuasion
Martino Bernasconi, Matteo Castiglioni, Andrea Celli, Alberto Marchesi, Francesc
o Trovò, Nicola Gatti
Bayesian persuasion studies how an informed sender should influence beliefs of r
ational receivers that take decisions through Bayesian updating of a common prio
r. We focus on the online Bayesian persuasion framework, in which the sender rep
eatedly faces one or more receivers with unknown and adversarially selected type
s. First, we show how to obtain a tight $\tilde O(T^{1/2})$ regret bound in the
case in which the sender faces a single receiver and has bandit feedback, improv
ing over the best previously known bound of $\tilde O(T^{4/5})$. Then, we provid
e the first no-regret guarantees for the multi-receiver setting under bandit fee
dback. Finally, we show how to design no-regret algorithms with polynomial per-i
teration running time by exploiting type reporting, thereby circumventing known
complexity results on online Bayesian persuasion. We provide efficient algorithm
s guaranteeing a $O(T^{1/2})$ regret upper bound both in the single- and multi-r
eceiver scenario when type reporting is allowed.
****

Constrained Phi-Equilibria
Martino Bernasconi, Matteo Castiglioni, Alberto Marchesi, Francesco Trovò, Nicol
a Gatti
The computational study of equilibria involving constraints on players' strategi
es has been largely neglected. However, in real-world applications, players are
usually subject to constraints ruling out the feasibility of some of their strat
egies, such as, e.g., safety requirements and budget caps. Computational studies
 on constrained versions of the Nash equilibrium have lead to some results under
 very stringent assumptions, while finding constrained versions of the correlate
d equilibrium (CE) is still unexplored. In this paper, we introduce and computat
ionally characterize constrained Phi-equilibria—a more general notion than const
rained CEs—in normal-form games. We show that computing such equilibria is in ge
neral computationally intractable, and also that the set of the equilibria may n
ot be convex, providing a sharp divide with unconstrained CEs. Nevertheless, we
provide a polynomial-time algorithm for computing a constrained (approximate) Ph
i-equilibrium maximizing a given linear function, when either the number of cons
traints or that of players' actions is fixed. Moreover, in the special case in w
hich a player's constraints do not depend on other players' strategies, we show
that an exact, function-maximizing equilibrium can be computed in polynomial tim
e, while one (approximate) equilibrium can be found with an efficient decentrali
zed no-regret learning algorithm.
****

Differentiable and Transportable Structure Learning
Jeroen Berrevoets, Nabeel Seedat, Fergus Imrie, Mihaela Van Der Schaar
Directed acyclic graphs (DAGs) encode a lot of information about a particular di
stribution in their structure. However, compute required to infer these structur
es is typically super-exponential in the number of variables, as inference requi
res a sweep of a combinatorially large space of potential structures. That is, u
ntil recent advances made it possible to search this space using a differentiabl
e metric, drastically reducing search time. While this technique— named NOTEARS
—is widely considered a seminal work in DAG-discovery, it concedes an important
property in favour of differentiability: transportability. To be transportable,
the structures discovered on one dataset must apply to another dataset from the
same domain. We introduce D-Struct which recovers transportability in the discov

ered structures through a novel architecture and loss function while remaining f
ully differentiable. Because D-Struct remains differentiable, our method can be
easily adopted in existing differentiable architectures, as was previously done
with NOTEARS. In our experiments, we empirically validate D-Struct with respect
to edge accuracy and structural Hamming distance in a variety of settings.
****

Polyhedral Complex Extraction from ReLU Networks using Edge Subdivision
Arturs Berzins
A neural network consisting of piecewise affine building blocks, such as fully-c
onnected layers and ReLU activations, is itself a piecewise affine function supp
orted on a polyhedral complex. This complex has been previously studied to chara
cterize theoretical properties of neural networks, but, in practice, extracting
it remains a challenge due to its high combinatorial complexity. A natural idea
described in previous works is to subdivide the regions via intersections with h
yperplanes induced by each neuron. However, we argue that this view leads to com
putational redundancy. Instead of regions, we propose to subdivide edges, leadin
g to a novel method for polyhedral complex extraction. A key to this are sign-ve
ctors, which encode the combinatorial structure of the complex. Our approach all
ows to use standard tensor operations on a GPU, taking seconds for millions of c
ells on a consumer grade machine. Motivated by the growing interest in neural sh
ape representation, we use the speed and differentiablility of our method to opt
imize geometric properties of the complex. The code is available at https://gith
ub.com/arturs-berzins/relu_edge_subdivision.
****

Robust One-Class Classification with Signed Distance Function using 1-Lipschitz
Neural Networks
Louis Béthune, Paul Novello, Guillaume Coiffier, Thibaut Boissin, Mathieu Serrur
ier, Quentin Vincenot, Andres Troya-Galvis
We propose a new method, dubbed One Class Signed Distance Function (OCSDF), to p
erform One Class Classification (OCC) by provably learning the Signed Distance F
unction (SDF) to the boundary of the support of any distribution. The distance t
o the support can be interpreted as a normality score, and its approximation usi
ng 1-Lipschitz neural networks provides robustness bounds against $l2$ adversari
al attacks, an under-explored weakness of deep learning-based OCC algorithms. As
 a result, OCSDF comes with a new metric, certified AUROC, that can be computed
at the same cost as any classical AUROC. We show that OCSDF is competitive again
st concurrent methods on tabular and image data while being way more robust to a
dversarial attacks, illustrating its theoretical properties. Finally, as explora
tory research perspectives, we theoretically and empirically show how OCSDF conn
ects OCC with image generation and implicit neural surface parametrization.
****

Neural Algorithmic Reasoning with Causal Regularisation
Beatrice Bevilacqua, Kyriacos Nikiforou, Borja Ibarz, Ioana Bica, Michela Pagani
ni, Charles Blundell, Jovana Mitrovic, Petar Veli■kovi■
Recent work on neural algorithmic reasoning has investigated the reasoning capab
ilities of neural networks, effectively demonstrating they can learn to execute
classical algorithms on unseen data coming from the train distribution. However,
 the performance of existing neural reasoners significantly degrades on out-of-d
istribution (OOD) test data, where inputs have larger sizes. In this work, we ma
ke an important observation: there are many different inputs for which an algori
thm will perform certain intermediate computations identically. This insight all
ows us to develop data augmentation procedures that, given an algorithm’s interm
ediate trajectory, produce inputs for which the target algorithm would have exac
tly the same next trajectory step. We ensure invariance in the next-step predict
ion across such inputs, by employing a self-supervised objective derived by our
observation, formalised in a causal graph. We prove that the resulting method, w
hich we call Hint-ReLIC, improves the OOD generalisation capabilities of the rea
soner. We evaluate our method on the CLRS algorithmic reasoning benchmark, where
 we show up to 3x improvements on the OOD test data.
****

Optimally-weighted Estimators of the Maximum Mean Discrepancy for Likelihood-Free Inference
Ayush Bharti, Masha Naslidnyk, Oscar Key, Samuel Kaski, Francois-Xavier Briol
Likelihood-free inference methods typically make use of a distance between simulated and real data. A common example is the maximum mean discrepancy (MMD), which has previously been used for approximate Bayesian computation, minimum distance estimation, generalised Bayesian inference, and within the nonparametric learning framework. The MMD is commonly estimated at a root-m rate, where m is the number of simulated samples. This can lead to significant computational challenges since a large m is required to obtain an accurate estimate, which is crucial for parameter estimation. In this paper, we propose a novel estimator for the MMD with significantly improved sample complexity. The estimator is particularly well suited for computationally expensive smooth simulators with low- to mid-dimensional inputs. This claim is supported through both theoretical results and an extensive simulation study on benchmark simulators.
****
Bandit Online Linear Optimization with Hints and Queries
Aditya Bhaskara, Ashok Cutkosky, Ravi Kumar, Manish Purohit
We study variants of the online linear optimization (OLO) problem with bandit feedback, where the algorithm has access to external information about the unknown cost vector. Our motivation is the recent body of work on using such “hints” towards improving regret bounds for OLO problems in the full-information setting. Unlike in the full-information OLO setting, with bandit feedback, we first show that one cannot improve the standard regret bounds of $\tilde{O}(\sqrt{T})$ by using hints, even if they are always well-correlated with the cost vector. In contrast, if the algorithm is empowered to issue queries and if all the responses are correct, then we show $O(\log T)$ regret is achievable. We then show how to make this result more robust—when some of the query responses can be adversarial—by using a little feedback on the quality of the responses.
****
Improved Online Conformal Prediction via Strongly Adaptive Online Learning
Aadyot Bhatnagar, Huan Wang, Caiming Xiong, Yu Bai
We study the problem of uncertainty quantification via prediction sets, in an online setting where the data distribution may vary arbitrarily over time. Recent work develops online conformal prediction techniques that leverage regret minimization algorithms from the online learning literature to learn prediction sets with approximately valid coverage and small regret. However, standard regret minimization is insufficient for handling changing environments, where performance guarantees may be desired not only over the full time horizon but also in all (sub-)intervals of time. We develop new online conformal prediction methods that minimize the strongly adaptive regret, which measures the worst-case regret over all intervals of a fixed length. We prove that our methods achieve near-optimal strongly adaptive regret for all interval lengths simultaneously, and approximately valid coverage. Experiments show that our methods consistently obtain better coverage and smaller prediction sets than existing methods on real-world tasks such as time series forecasting and image classification under distribution shift.
****
Data-Copying in Generative Models: A Formal Framework
Robi Bhattacharjee, Sanjoy Dasgupta, Kamalika Chaudhuri
There has been some recent interest in detecting and addressing memorization of training data by deep neural networks. A formal framework for memorization in generative models, called “data-copying” was proposed by Meehan et. al (2020). We build upon their work to show that their framework may fail to detect certain kinds of blatant memorization. Motivated by this and the theory of non-parametric methods, we provide an alternative definition of data-copying that applies more locally. We provide a method to detect data-copying, and provably show that it works with high probability when enough data is available. We also provide lower bounds that characterize the sample requirement for reliable detection.
****

Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling
Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O’Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, Oskar Van Der Wal
How do large language models (LLMs) develop and evolve over the course of training? How do these patterns change as models scale? To answer these questions, we introduce Pythia, a suite of 16 LLMs all trained on public data seen in the exact same order and ranging in size from 70M to 12B parameters. We provide public access to 154 checkpoints for each one of the 16 models, alongside tools to download and reconstruct their exact training dataloaders for further study. We intend Pythia to facilitate research in many areas, and we present several case studies including novel results in memorization, term frequency effects on few-shot performance, and reducing gender bias. We demonstrate that this highly controlled setup can be used to yield novel insights toward LLMs and their training dynamics. Trained models, analysis code, training code, and training data can be found at https://github.com/EleutherAI/pythia.
****

StriderNet: A Graph Reinforcement Learning Approach to Optimize Atomic Structures on Rough Energy Landscapes
Vaibhav Bihani, Sahil Manchanda, Srikanth Sastry, Sayan Ranu, N M Anoop Krishnan
Optimization of atomic structures presents a challenging problem, due to their highly rough and non-convex energy landscape, with wide applications in the fields of drug design, materials discovery, and mechanics. Here, we present a graph reinforcement learning approach, StriderNet, that learns a policy to displace the atoms towards low energy configurations. We evaluate the performance of StriderNet on three complex atomic systems, namely, binary Lennard-Jones particles, calcium silicate hydrates gel, and disordered silicon. We show that StriderNet outperforms all classical optimization algorithms and enables the discovery of a lower energy minimum. In addition, StriderNet exhibits a higher rate of reaching minima with energies, as confirmed by the average over multiple realizations. Finally, we show that StriderNet exhibits inductivity to unseen system sizes that are an order of magnitude different from the training system. All the codes and datasets are available at https://github.com/M3RG-IITD/StriderNET.
****

Modeling Temporal Data as Continuous Functions with Stochastic Process Diffusion
Marin Biloš, Kashif Rasul, Anderson Schneider, Yuriy Nevmyvaka, Stephan Günneman
n
Temporal data such as time series can be viewed as discretized measurements of the underlying function. To build a generative model for such data we have to model the stochastic process that governs it. We propose a solution by defining the denoising diffusion model in the function space which also allows us to naturally handle irregularly-sampled observations. The forward process gradually adds noise to functions, preserving their continuity, while the learned reverse process removes the noise and returns functions as new samples. To this end, we define suitable noise sources and introduce novel denoising and score-matching models. We show how our method can be used for multivariate probabilistic forecasting and imputation, and how our model can be interpreted as a neural process.
****

In or Out? Fixing ImageNet Out-of-Distribution Detection Evaluation
Julian Bitterwolf, Maximilian Müller, Matthias Hein
Out-of-distribution (OOD) detection is the problem of identifying inputs which are unrelated to the in-distribution task. The OOD detection performance when the in-distribution (ID) is ImageNet-1K is commonly being tested on a small range of test OOD datasets. We find that most of the currently used test OOD datasets, including datasets from the open set recognition (OSR) literature, have severe issues: In some cases more than 50$%$ of the dataset contains objects belonging to one of the ID classes. These erroneous samples heavily distort the evaluation of OOD detectors. As a solution, we introduce with NINCO a novel test OOD dataset, each sample checked to be ID free, which with its fine-grained range of OOD classes allows for a detailed analysis of an OOD detector’s strengths and failure

modes, particularly when paired with a number of synthetic “OOD unit-tests”. We provide detailed evaluations across a large set of architectures and OOD detection methods on NINCO and the unit-tests, revealing new insights about model weaknesses and the effects of pretraining on OOD detection performance. We provide code and data at https://github.com/j-cb/NINCO.

****

Invariant Slot Attention: Object Discovery with Slot-Centric Reference Frames

Ondrej Biza, Sjoerd Van Steenkiste, Mehdi S. M. Sajjadi, Gamaleldin Fathy Elsayed, Aravindh Mahendran, Thomas Kipf

Automatically discovering composable abstractions from raw perceptual data is a long-standing challenge in machine learning. Recent slot-based neural networks that learn about objects in a self-supervised manner have made exciting progress in this direction. However, they typically fall short at adequately capturing spatial symmetries present in the visual world, which leads to sample inefficiency, such as when entangling object appearance and pose. In this paper, we present a simple yet highly effective method for incorporating spatial symmetries via slot-centric reference frames. We incorporate equivariance to per-object pose transformations into the attention and generation mechanism of Slot Attention by translating, scaling, and rotating position encodings. These changes result in little computational overhead, are easy to implement, and can result in large gains in terms of data efficiency and overall improvements to object discovery. We evaluate our method on a wide range of synthetic object discovery benchmarks namely CLEVR, Tetrominoes, CLEVRTex, Objects Room and MultiShapeNet, and show promising improvements on the challenging real-world Waymo Open dataset.

****

Understanding Oversquashing in GNNs through the Lens of Effective Resistance

Mitchell Black, Zhengchao Wan, Amir Nayyeri, Yusu Wang

Message passing graph neural networks (GNNs) are a popular learning architectures for graph-structured data. However, one problem GNNs experience is oversquashing, where a GNN has difficulty sending information between distant nodes. Understanding and mitigating oversquashing has recently received significant attention from the research community. In this paper, we continue this line of work by analyzing oversquashing through the lens of the effective resistance between nodes in the input graph. Effective resistance intuitively captures the “strength” of connection between two nodes by paths in the graph, and has a rich literature spanning many areas of graph theory. We propose to use total effective resistance as a bound of the total amount of oversquashing in a graph and provide theoretical justification for its use. We further develop an algorithm to identify edges to be added to an input graph to minimize the total effective resistance, thereby alleviating oversquashing. We provide empirical evidence of the effectiveness of our total effective resistance based rewiring strategies for improving the performance of GNNs.

****

Unit Scaling: Out-of-the-Box Low-Precision Training

Charlie Blake, Douglas Orr, Carlo Luschi

We present unit scaling, a paradigm for designing deep learning models that simplifies the use of low-precision number formats. Training in FP16 or the recently proposed FP8 formats offers substantial efficiency gains, but can lack sufficient range for out-of-the-box training. Unit scaling addresses this by introducing a principled approach to model numerics: seeking unit variance of all weights, activations and gradients at initialisation. Unlike alternative methods, this approach neither requires multiple training runs to find a suitable scale nor has significant computational overhead. We demonstrate the efficacy of unit scaling across a range of models and optimisers. We further show that existing models can be adapted to be unit-scaled, training BERT-Large in FP16 and then FP8 with no degradation in accuracy.

****

FLEX: an Adaptive Exploration Algorithm for Nonlinear Systems

Matthieu Blanke, Marc Lelarge

Model-based reinforcement learning is a powerful tool, but collecting data to fi

t an accurate model of the system can be costly. Exploring an unknown environmen t in a sample-efficient manner is hence of great importance. However, the comple xity of dynamics and the computational limitations of real systems make this tas k challenging. In this work, we introduce FLEX, an exploration algorithm for non linear dynamics based on optimal experimental design. Our policy maximizes the i nformation of the next step and results in an adaptive exploration algorithm, co mpatible with arbitrary parametric learning models, and requiring minimal comput ing resources. We test our method on a number of nonlinear environments covering different settings, including time-varying dynamics. Keeping in mind that explo ration is intended to serve an exploitation objective, we also test our algorith m on downstream model-based classical control tasks and compare it to other stat e-of-the-art model-based and model-free approaches. The performance achieved by FLEX is competitive and its computational cost is low.

****

Not all Strongly Rayleigh Distributions Have Small Probabilistic Generating Circ uits

Markus Bläser

Probabilistic modeling is a central task in machine learning. Probabilistic mode ls should be tractable, i.e., allowing tractable probabilistic inference, but al so efficient, i.e., being able to represent a large set of probability distribut ions. Zhang et al. (ICML 2021) recently proposed a new model, probabilistic gene rating circuits. They raised the question whether every strongly Rayleigh distri bution can be efficiently represented by such circuits. We prove that this quest ion has a negative answer, there are strongly Rayleigh distributions that cannot be represented by polynomial-sized probabilistic generating circuits, assuming a widely accepted complexity theoretic conjecture.

****

Learning the Dynamics of Sparsely Observed Interacting Systems

Linus Bleistein, Adeline Fermanian, Anne-Sophie Jannot, Agathe Guilloux

We address the problem of learning the dynamics of an unknown non-parametric sys tem linking a target and a feature time series. The feature time series is measu red on a sparse and irregular grid, while we have access to only a few points of the target time series. Once learned, we can use these dynamics to predict valu es of the target from the previous values of the feature time series. We frame t his task as learning the solution map of a controlled differential equation (CDE). By leveraging the rich theory of signatures, we are able to cast this non-lin ear problem as a high-dimensional linear regression. We provide an oracle bound on the prediction error which exhibits explicit dependencies on the individual-s pecific sampling schemes. Our theoretical results are illustrated by simulations which show that our method outperforms existing algorithms for recovering the f ull time series while being computationally cheap. We conclude by demonstrating its potential on real-world epidemiological data.

****

Subset Selection Based On Multiple Rankings in the Presence of Bias: Effectivene ss of Fairness Constraints for Multiwinner Voting Score Functions

Niclas Boehmer, L. Elisa Celis, Lingxiao Huang, Anay Mehrotra, Nisheeth K. Vishn oi

We consider the problem of subset selection where one is given multiple rankings of items and the goal is to select the highest "quality" subset. Score function s from the multiwinner voting literature have been used to aggregate rankings in to quality scores for subsets. We study this setting of subset selection problem s when, in addition, rankings may contain systemic or unconscious biases toward a group of items. For a general model of input rankings and biases, we show that requiring the selected subset to satisfy group fairness constraints can improve the quality of the selection with respect to unbiased rankings. Importantly, we show that for fairness constraints to be effective, different multiwinner score functions may require a drastically different number of rankings: While for som e functions, fairness constraints need an exponential number of rankings to reco ver a close-to-optimal solution, for others, this dependency is only polynomial. This result relies on a novel notion of "smoothness" of submodular functions in

this setting that quantifies how well a function can "correctly" assess the quality of items in the presence of bias. The results in this paper can be used to guide the choice of multiwinner score functions for the subset selection setting considered here; we additionally provide a tool to empirically enable this.

****

Properties of the Mallows Model Depending on the Number of Alternatives: A Warning for an Experimentalist

Niclas Boehmer, Piotr Faliszewski, Sonja Kraiczy

The Mallows model is a popular distribution for ranked data. We empirically and theoretically analyze how the properties of rankings sampled from the Mallows model change when increasing the number of alternatives. We find that real-world data behaves differently from the Mallows model, yet is in line with its recent variant proposed by Boehmer et al. [IJCAI '21]. As part of our study, we issue several warnings about using the classic Mallows model. For instance, we find that one should be extremely careful when using the Mallows model to generate data for experiments with a varying number of alternatives, as observed trends in such experiments might be due to the changing nature of the generated data.

****

A Robust Optimisation Perspective on Counterexample-Guided Repair of Neural Networks

David Boetius, Stefan Leue, Tobias Sutter

Counterexample-guided repair aims at creating neural networks with mathematical safety guarantees, facilitating the application of neural networks in safety-critical domains. However, whether counterexample-guided repair is guaranteed to terminate remains an open question. We approach this question by showing that counterexample-guided repair can be viewed as a robust optimisation algorithm. While termination guarantees for neural network repair itself remain beyond our reach, we prove termination for more restrained machine learning models and disprove termination in a general setting. We empirically study the practical implications of our theoretical results, demonstrating the suitability of common verifiers and falsifiers for repair despite a disadvantageous theoretical result. Additionally, we use our theoretical insights to devise a novel algorithm for repairing linear regression models based on quadratic programming, surpassing existing approaches.

****

Beyond the Universal Law of Robustness: Sharper Laws for Random Features and Neural Tangent Kernels

Simone Bombari, Shayan Kiyani, Marco Mondelli

Machine learning models are vulnerable to adversarial perturbations, and a thought-provoking paper by Bubeck and Sellke has analyzed this phenomenon through the lens of over-parameterization: interpolating smoothly the data requires significantly more parameters than simply memorizing it. However, this "universal" law provides only a necessary condition for robustness, and it is unable to discriminate between models. In this paper, we address these gaps by focusing on empirical risk minimization in two prototypical settings, namely, random features and the neural tangent kernel (NTK). We prove that, for random features, the model is not robust for any degree of over-parameterization, even when the necessary condition coming from the universal law of robustness is satisfied. In contrast, for even activations, the NTK model meets the universal lower bound, and it is robust as soon as the necessary condition on over-parameterization is fulfilled. This also addresses a conjecture in prior work by Bubeck, Li and Nagaraj. Our analysis decouples the effect of the kernel of the model from an "interaction matrix", which describes the interaction with the test data and captures the effect of the activation. Our theoretical results are corroborated by numerical evidence on both synthetic and standard datasets (MNIST, CIFAR-10).

****

Sliced-Wasserstein on Symmetric Positive Definite Matrices for M/EEG Signals

Clément Bonet, Beno■■t Malézieux, Alain Rakotomamonjy, Lucas Drumetz, Thomas Moreau, Matthieu Kowalski, Nicolas Courty

When dealing with electro or magnetoencephalography records, many supervised pre

diction tasks are solved by working with covariance matrices to summarize the si
gnals. Learning with these matrices requires the usage of Riemanian geometry to
account for their structure. In this paper, we propose a new method to deal with
 distributions of covariance matrices, and demonstrate its computational efficie
ncy on M/EEG multivariate time series. More specifically, we define a Sliced-Was
serstein distance between measures of symmetric positive definite matrices that
comes with strong theoretical guarantees. Then, we take advantage of its propert
ies and kernel methods to apply this discrepancy to brain-age prediction from ME
G data, and compare it to state-of-the-art algorithms based on Riemannian geomet
ry. Finally, we show that it is an efficient surrogate to the Wasserstein distan
ce in domain adaptation for Brain Computer Interface applications.
****

Spherical Fourier Neural Operators: Learning Stable Dynamics on the Sphere
Boris Bonev, Thorsten Kurth, Christian Hundt, Jaideep Pathak, Maximilian Baust,
Karthik Kashinath, Anima Anandkumar
Fourier Neural Operators (FNOs) have proven to be an efficient and effective met
hod for resolution-independent operator learning in a broad variety of applicati
on areas across scientific machine learning. A key reason for their success is t
heir ability to accurately model long-range dependencies in spatio-temporal data
 by learning global convolutions in a computationally efficient manner. To this
end, FNOs rely on the discrete Fourier transform (DFT), however, DFTs cause visu
al and spectral artifacts as well as pronounced dissipation when learning operat
ors in spherical coordinates by incorrectly assuming flat geometry. To overcome
this limitation, we generalize FNOs on the sphere, introducing Spherical FNOs (S
FNOs) for learning operators on spherical geometries. We apply SFNOs to forecast
ing atmo- spheric dynamics, and demonstrate stable autoregressive rollouts for a
 year of simulated time (1,460 steps), while retaining physically plausible dyna
mics. The SFNO has important implications for machine learning-based simulation
of climate dynamics that could eventually help accelerate our response to climat
e change.
****

The Regret of Exploration and the Control of Bad Episodes in Reinforcement Learn
ing
Victor Boone, Bruno Gaujal
The first contribution of this paper is the introduction of a new performance me
asure of a RL algorithm that is more discriminating than the regret, that we cal
l the regret of exploration that measures the asymptotic cost of exploration. Th
e second contribution is a new performance test (PT) to end episodes in RL optim
istic algorithms. This test is based on the performance of the current policy wi
th respect to the best policy over the current confidence set. This is in contra
st with all existing RL algorithms whose episode lengths are only based on the n
umber of visits to the states. This modification does not harm the regret and br
ings an additional property. We show that while all current episodic RL algorith
ms have a linear regret of exploration, our method has a $O(\log{T})$ regret of
exploration for non-degenerate deterministic MDPs.
****

Model-agnostic Measure of Generalization Difficulty
Akhilan Boopathy, Kevin Liu, Jaedong Hwang, Shu Ge, Asaad Mohammedsaleh, Ila R F
iete
The measure of a machine learning algorithm is the difficulty of the tasks it ca
n perform, and sufficiently difficult tasks are critical drivers of strong machi
ne learning models. However, quantifying the generalization difficulty of machin
e learning benchmarks has remained challenging. We propose what is to our knowle
dge the first model-agnostic measure of the inherent generalization difficulty o
f tasks. Our inductive bias complexity measure quantifies the total information
required to generalize well on a task minus the information provided by the data
. It does so by measuring the fractional volume occupied by hypotheses that gene
ralize on a task given that they fit the training data. It scales exponentially
with the intrinsic dimensionality of the space over which the model must general
ize but only polynomially in resolution per dimension, showing that tasks which

require generalizing over many dimensions are drastically more difficult than ta
sks involving more detail in fewer dimensions. Our measure can be applied to com
pute and compare supervised learning, reinforcement learning and meta-learning g
eneralization difficulties against each other. We show that applied empirically,
 it formally quantifies intuitively expected trends, e.g. that in terms of requi
red inductive bias, MNIST $<$ CIFAR10 $<$ Imagenet and fully observable Markov d
ecision processes (MDPs) $<$ partially observable MDPs. Further, we show that cl
assification of complex images $<$ few-shot meta-learning with simple images. Ou
r measure provides a quantitative metric to guide the construction of more compl
ex tasks requiring greater inductive bias, and thereby encourages the developmen
t of more sophisticated architectures and learning algorithms with more powerful
 generalization capabilities.
****

Returning The Favour: When Regression Benefits From Probabilistic Causal Knowled
ge

Shahine Bouabid, Jake Fawkes, Dino Sejdinovic

A directed acyclic graph (DAG) provides valuable prior knowledge that is often d
iscarded in regression tasks in machine learning. We show that the independences
 arising from the presence of collider structures in DAGs provide meaningful ind
uctive biases, which constrain the regression hypothesis space and improve predi
ctive performance. We introduce collider regression, a framework to incorporate
probabilistic causal knowledge from a collider in a regression problem. When the
 hypothesis space is a reproducing kernel Hilbert space, we prove a strictly pos
itive generalisation benefit under mild assumptions and provide closed-form esti
mators of the empirical risk minimiser. Experiments on synthetic and climate mod
el data demonstrate performance gains of the proposed methodology.
****

In Search for a Generalizable Method for Source Free Domain Adaptation

Malik Boudiaf, Tom Denton, Bart Van Merrienboer, Vincent Dumoulin, Eleni Trianta
fillou

Source-free domain adaptation (SFDA) is compelling because it allows adapting an
 off-the-shelf model to a new domain using only unlabelled data. In this work, w
e apply existing SFDA techniques to a challenging set of naturally-occurring dis
tribution shifts in bioacoustics, which are very different from the ones commonl
y studied in computer vision. We find existing methods perform differently relat
ive to each other than observed in vision benchmarks, and sometimes perform wors
e than no adaptation at all. We propose a new simple method which outperforms th
e existing methods on our new shifts while exhibiting strong performance on a ra
nge of vision datasets. Our findings suggest that existing SFDA methods are not
as generalizable as previously thought and that considering diverse modalities c
an be a useful avenue for designing more robust models.
****

Quantum Speedups for Zero-Sum Games via Improved Dynamic Gibbs Sampling

Adam Bouland, Yosheb M Getachew, Yujia Jin, Aaron Sidford, Kevin Tian

We give a quantum algorithm for computing an ϵ-approximate Nash equilib
rium of a zero-sum game in a $m \times n$ payoff matrix with bounded entries. Gi
ven a standard quantum oracle for accessing the payoff matrix our algorithm runs
 in time $\widetilde{O}(\sqrt{m + n}\cdot \epsilon^{-2.5} + \epsilon^{-3})$ and
outputs a classical representation of the ϵ-approximate Nash equilibriu
m. This improves upon the best prior quantum runtime of $\widetilde{O}(\sqrt{m +
 n} \cdot \epsilon^{-3})$ obtained by [van Apeldoorn, Gilyen ’19] and the classi
cal $\widetilde{O}((m + n) \cdot \epsilon^{-2})$ runtime due to [Grigoradis, Kha
chiyan ’95] whenever $\epsilon = \Omega((m +n)^{-1})$. We obtain this result by
designing new quantum data structures for efficiently sampling from a slowly-cha
nging Gibbs distribution.
****

Diffusion Models as Artists: Are we Closing the Gap between Humans and Machines?

Victor Boutin, Thomas Fel, Lakshya Singhal, Rishav Mukherji, Akash Nagaraj, Juli
en Colin, Thomas Serre

An important milestone for AI is the development of algorithms that can produce

drawings that are indistinguishable from those of humans. Here, we adapt the ”diversity vs. recognizability” scoring framework from Boutin et al (2022) and find that one-shot diffusion models have indeed started to close the gap between humans and machines. However, using a finer-grained measure of the originality of individual samples, we show that strengthening the guidance of diffusion models helps improve the humanness of their drawings, but they still fall short of approximating the originality and recognizability of human drawings. Comparing human category diagnostic features, collected through an online psychophysics experiment, against those derived from diffusion models reveals that humans rely on fewer and more localized features. Overall, our study suggests that diffusion models have significantly helped improve the quality of machine-generated drawings; however, a gap between humans and machines remains – in part explainable by discrepancies in visual strategies.

****

Settling the Reward Hypothesis

Michael Bowling, John D Martin, David Abel, Will Dabney

The reward hypothesis posits that, "all of what we mean by goals and purposes can be well thought of as maximization of the expected value of the cumulative sum of a received scalar signal (reward)." We aim to fully settle this hypothesis. This will not conclude with a simple affirmation or refutation, but rather specify completely the implicit requirements on goals and purposes under which the hypothesis holds.

****

ILLUME: Rationalizing Vision-Language Models through Human Interactions

Manuel Brack, Patrick Schramowski, Björn Deiseroth, Kristian Kersting

Bootstrapping from pre-trained language models has been proven to be an efficient approach for building vision-language models (VLM) for tasks such as image captioning or visual question answering. However, outputs of these models rarely align with user’s rationales for specific answers. In order to improve this alignment and reinforce commonsense reasons, we propose a tuning paradigm based on human interactions with machine-generated data. Our ILLUME executes the following loop: Given an image-question-answer prompt, the VLM samples multiple candidate rationales, and a human critic provides feedback via preference selection, used for fine-tuning. This loop increases the training data and gradually carves out the VLM’s rationalization capabilities that are aligned with human intent. Our exhaustive experiments demonstrate that ILLUME is competitive with standard supervised finetuning while using significantly fewer training data and only requiring minimal feedback.

****

Provably Learning Object-Centric Representations

Jack Brady, Roland S. Zimmermann, Yash Sharma, Bernhard Schölkopf, Julius Von Kügelgen, Wieland Brendel

Learning structured representations of the visual world in terms of objects promises to significantly improve the generalization abilities of current machine learning models. While recent efforts to this end have shown promising empirical progress, a theoretical account of when unsupervised object-centric representation learning is possible is still lacking. Consequently, understanding the reasons for the success of existing object-centric methods as well as designing new theoretically grounded methods remains challenging. In the present work, we analyze when object-centric representations can provably be learned without supervision. To this end, we first introduce two assumptions on the generative process for scenes comprised of several objects, which we call compositionality and irreducibility. Under this generative process, we prove that the ground-truth object representations can be identified by an invertible and compositional inference model, even in the presence of dependencies between objects. We empirically validate our results through experiments on synthetic data. Finally, we provide evidence that our theory holds predictive power for existing object-centric models by showing a close correspondence between models’ compositionality and invertibility and their empirical identifiability.

****

Quantifying Human Priors over Social and Navigation Networks
Gecia Bravo-Hermsdorff
Human knowledge is largely implicit and relational — do we have a friend in comm on? can I walk from here to there? In this work, we leverage the combinatorial s tructure of graphs to quantify human priors over such relational data. Our exper iments focus on two domains that have been continuously relevant over evolutiona ry timescales: social interaction and spatial navigation. We find that some feat ures of the inferred priors are remarkably consistent, such as the tendency for sparsity as a function of graph size. Other features are domain-specific, such a s the propensity for triadic closure in social interactions. More broadly, our w ork demonstrates how nonclassical statistical analysis of indirect behavioral ex periments can be used to efficiently model latent biases in the data.
****
Critical Points and Convergence Analysis of Generative Deep Linear Networks Trai ned with Bures-Wasserstein Loss
Pierre Bréchet, Katerina Papagiannouli, Jing An, Guido Montufar
We consider a deep matrix factorization model of covariance matrices trained wit h the Bures-Wasserstein distance. While recent works have made advances in the s tudy of the optimization problem for overparametrized low-rank matrix approximat ion, much emphasis has been placed on discriminative settings and the square los s. In contrast, our model considers another type of loss and connects with the g enerative setting. We characterize the critical points and minimizers of the Bur es-Wasserstein distance over the space of rank-bounded matrices. The Hessian of this loss at low-rank matrices can theoretically blow up, which creates challeng es to analyze convergence of gradient optimization methods. We establish converg ence results for gradient flow using a smooth perturbative version of the loss a s well as convergence results for finite step size gradient descent under certai n assumptions on the initial weights.
****
Emergence of Sparse Representations from Noise
Trenton Bricken, Rylan Schaeffer, Bruno Olshausen, Gabriel Kreiman
A hallmark of biological neural networks, which distinguishes them from their ar tificial counterparts, is the high degree of sparsity in their activations. This discrepancy raises three questions our work helps to answer: (i) Why are biolog ical networks so sparse? (ii) What are the benefits of this sparsity? (iii) How can these benefits be utilized by deep learning models? Our answers to all of th ese questions center around training networks to handle random noise. Surprising ly, we discover that noisy training introduces three implicit loss terms that re sult in sparsely firing neurons specializing to high variance features of the da taset. When trained to reconstruct noisy-CIFAR10, neurons learn biological recep tive fields. More broadly, noisy training presents a new approach to potentially increase model interpretability with additional benefits to robustness and comp utational efficiency.
****
Differentially Private Optimization on Large Model at Small Cost
Zhiqi Bu, Yu-Xiang Wang, Sheng Zha, George Karypis
Differentially private (DP) optimization is the standard paradigm to learn large neural networks that are accurate and privacy-preserving. The computational cos t for DP deep learning, however, is notoriously heavy due to the per-sample grad ient clipping. Existing DP implementations are 2$\sim$1000$\times$ more costly i n time and space complexity than the standard (non-private) training. In this wo rk, we develop a novel Book-Keeping (BK) technique that implements existing DP o ptimizers (thus achieving the same accuracy), with a substantial improvement on the computational cost. Specifically, BK enables DP training on large models and high dimensional data to be roughly as fast and memory-saving as the standard t raining, whereas previous DP algorithms can be inefficient or incapable of train ing due to memory error. The computational advantage of BK is supported by the c omplexity analysis as well as extensive experiments on vision and language tasks . Our implementation achieves state-of-the-art (SOTA) accuracy with very small e xtra cost: on GPT2 and at almost the same memory cost ($<$1% overhead), BK has 1

.03$\times$ the time complexity of the standard training (0.83$\times$ training speed in practice), and 0.61$\times$ the time complexity of the most efficient DP implementation (1.36$\times$ training speed in practice). We open-source the codebase for the BK algorithm at https://github.com/awslabs/fast-differential-privacy.

****

Machine Learning Force Fields with Data Cost Aware Training

Alexander Bukharin, Tianyi Liu, Shengjie Wang, Simiao Zuo, Weihao Gao, Wen Yan, Tuo Zhao

Machine learning force fields (MLFF) have been proposed to accelerate molecular dynamics (MD) simulation, which finds widespread applications in chemistry and biomedical research. Even for the most data-efficient MLFFs, reaching chemical accuracy can require hundreds of frames of force and energy labels generated by expensive quantum mechanical algorithms, which may scale as $O(n^3)$ to $O(n^7)$, with n proportional to the number of basis functions. To address this issue, we propose a multi-stage computational framework – ASTEROID, which lowers the data cost of MLFFs by leveraging a combination of cheap inaccurate data and expensive accurate data. The motivation behind ASTEROID is that inaccurate data, though incurring large bias, can help capture the sophisticated structures of the underlying force field. Therefore, we first train a MLFF model on a large amount of inaccurate training data, employing a bias-aware loss function to prevent the model from overfitting the potential bias of this data. We then fine-tune the obtained model using a small amount of accurate training data, which preserves the knowledge learned from the inaccurate training data while significantly improving the model's accuracy. Moreover, we propose a variant of ASTEROID based on score matching for the setting where the inaccurate training data are unlabeled. Extensive experiments on MD datasets and downstream tasks validate the efficacy of ASTEROID. Our code and data are available at https://github.com/abukharin3/asteroid.

****

Label differential privacy and private training data release

Robert Istvan Busa-Fekete, Andres Munoz Medina, Umar Syed, Sergei Vassilvitskii

We study differentially private mechanisms for sharing training data in machine learning settings. Our goal is to enable learning of an accurate predictive model while protecting the privacy of each user's label. Previous work established privacy guarantees that assumed the features are public and given exogenously, a setting known as label differential privacy. In some scenarios, this can be a strong assumption that removes the interplay between features and labels from the privacy analysis. We relax this approach and instead assume the features are drawn from a distribution that depends on the private labels. We first show that simply adding noise to the label, as in previous work, can lead to an arbitrarily weak privacy guarantee, and also present methods for estimating this privacy loss from data. We then present a new mechanism that replaces some training examples with synthetically generated data, and show that our mechanism has a much better privacy-utility tradeoff if the synthetic data is 'realistic', in a certain quantifiable sense. Finally, we empirically validate our theoretical analysis.

****

The SSL Interplay: Augmentations, Inductive Bias, and Generalization

Vivien Cabannes, Bobak Kiani, Randall Balestriero, Yann Lecun, Alberto Bietti

Self-supervised learning (SSL) has emerged as a powerful framework to learn representations from raw data without supervision. Yet in practice, engineers face issues such as instability in tuning optimizers and collapse of representations during training. Such challenges motivate the need for a theory to shed light on the complex interplay between the choice of data augmentation, network architecture, and training algorithm. % on the resulting performance in downstream tasks. We study such an interplay with a precise analysis of generalization performance on both pretraining and downstream tasks in kernel regimes, and highlight several insights for SSL practitioners that arise from our theory.

****

Online Mechanism Design for Information Acquisition

Federico Cacciamani, Matteo Castiglioni, Nicola Gatti
We study the problem of designing mechanisms for information acquisition scenarios. This setting models strategic interactions between a uniformed receiver and a set of informed senders. In our model the senders receive information about the underlying state of nature and communicate their observation (either truthfully or not) to the receiver, which, based on this information, selects an action. Our goal is to design mechanisms maximizing the receiver's utility while incentivizing the senders to report truthfully their information. First, we provide an algorithm that efficiently computes an optimal incentive compatible (IC) mechanism. Then, we focus on the online problem in which the receiver sequentially interacts in an unknown game, with the objective of minimizing the cumulative regret w.r.t. the optimal IC mechanism, and the cumulative violation of the incentive compatibility constraints. We investigate two different online scenarios, i.e., the full and bandit feedback settings. For the full feedback problem, we propose an algorithm that guarantees $\tilde{O}(\sqrt{T})$ regret and violation, while for the bandit feedback setting we present an algorithm that attains $\tilde{O}(T^{\alpha})$ regret and $\tilde{O}(T^{1-\alpha/2})$ violation for any $\alpha \in [1/2, 1]$. Finally, we complement our results providing a tight lower bound.
****
MyoDex: A Generalizable Prior for Dexterous Manipulation
Vittorio Caggiano, Sudeep Dasari, Vikash Kumar
Human dexterity is a hallmark of motor control behaviors. Our hands can rapidly synthesize new behaviors despite the complexity (multi-articular and multi-joints, with 23 joints controlled by more than 40 muscles) of mosculoskeletal control. In this work, we take inspiration from how human dexterity builds on a diversity of prior experiences, instead of being acquired through a single task. Motivated by this observation, we set out to develop agents that can build upon previous experience to quickly acquire new (previously unattainable) behaviors. Specifically, our approach leverages multi-task learning to implicitly capture a task-agnostic behavioral priors (MyoDex) for human-like dexterity, using a physiologically realistic human hand model – MyoHand. We demonstrate MyoDex's effectiveness in few-shot generalization as well as positive transfer to a large repertoire of unseen dexterous manipulation tasks. MyoDex can solve approximately 3x more tasks and it can accelerate the achievement of solutions by about 4x in comparison to a distillation baseline. While prior work has synthesized single musculoskeletal control behaviors, MyoDex is the first generalizable manipulation prior that catalyzes the learning of dexterous physiological control across a large variety of contact-rich behaviors.
****
What Can Be Learnt With Wide Convolutional Neural Networks?
Francesco Cagnetta, Alessandro Favero, Matthieu Wyart
Understanding how convolutional neural networks (CNNs) can efficiently learn high-dimensional functions remains a fundamental challenge. A popular belief is that these models harness the local and hierarchical structure of natural data such as images. Yet, we lack a quantitative understanding of how such structure affects performance, e.g., the rate of decay of the generalisation error with the number of training samples. In this paper, we study infinitely-wide deep CNNs in the kernel regime. First, we show that the spectrum of the corresponding kernel inherits the hierarchical structure of the network, and we characterise its asymptotics. Then, we use this result together with generalisation bounds to prove that deep CNNs adapt to the spatial scale of the target function. In particular, we find that if the target function depends on low-dimensional subsets of adjacent input variables, then the decay of the error is controlled by the effective dimensionality of these subsets. Conversely, if the target function depends on the full set of input variables, then the error decay is controlled by the input dimension. We conclude by computing the generalisation error of a deep CNN trained on the output of another deep CNN with randomly-initialised parameters. Interestingly, we find that, despite their hierarchical structure, the functions generated by infinitely-wide deep CNNs are too rich to be efficiently learnable in high dimension.

****
Causal Discovery with Latent Confounders Based on Higher-Order Cumulants
Ruichu Cai, Zhiyi Huang, Wei Chen, Zhifeng Hao, Kun Zhang
Causal discovery with latent confounders is an important but challenging task in many scientific areas. Despite the success of some overcomplete independent component analysis (OICA) based methods in certain domains, they are computationally expensive and can easily get stuck into local optima. We notice that interestingly, by making use of higher-order cumulants, there exists a closed-form solution to OICA in specific cases, e.g., when the mixing procedure follows the One-Latent-Component structure. In light of the power of the closed-form solution to OICA corresponding to the One-Latent-Component structure, we formulate a way to estimate the mixing matrix using the higher-order cumulants, and further propose the testable One-Latent-Component condition to identify the latent variables and determine causal orders. By iteratively removing the share identified latent components, we successfully extend the results on the One-Latent-Component structure to the Multi-Latent-Component structure and finally provide a practical and asymptotically correct algorithm to learn the causal structure with latent variables. Experimental results illustrate the asymptotic correctness and effectiveness of the proposed method.
****
On the Connection Between MPNN and Graph Transformer
Chen Cai, Truong Son Hy, Rose Yu, Yusu Wang
Graph Transformer (GT) recently has emerged as a new paradigm of graph learning algorithms, outperforming the previously popular Message Passing Neural Network (MPNN) on multiple benchmarks. Previous work shows that with proper position embedding, GT can approximate MPNN arbitrarily well, implying that GT is at least as powerful as MPNN. In this paper, we study the inverse connection and show that MPNN with virtual node (VN), a commonly used heuristic with little theoretical understanding, is powerful enough to arbitrarily approximate the self-attention layer of GT. In particular, we first show that if we consider one type of linear transformer, the so-called Performer/Linear Transformer, then MPNN + VN with only $\mathcal{O}(1)$ depth and $\mathcal{O}(1)$ width can approximate a self-attention layer in Performer/Linear Transformer. Next, via a connection between MPNN + VN and DeepSets, we prove the MPNN + VN with $\mathcal{O}(n^d)$ width and $\mathcal{O}(1)$ depth can approximate the self-attention layer arbitrarily well, where d is the input feature dimension. Lastly, under some assumptions, we provide an explicit construction of MPNN + VN with $\mathcal{O}(1)$ width and $\mathcal{O}(n)$ depth approximating the self-attention layer in GT arbitrarily well. On the empirical side, we demonstrate that 1) MPNN + VN is a surprisingly strong baseline, outperforming GT on the recently proposed Long Range Graph Benchmark (LRGB) dataset, 2) our MPNN + VN improves over early implementation on a wide range of OGB datasets and 3) MPNN + VN outperforms Linear Transformer and MPNN on the climate modeling task.
****
Ske2Grid: Skeleton-to-Grid Representation Learning for Action Recognition
Dongqi Cai, Yangyuxuan Kang, Anbang Yao, Yurong Chen
This paper presents Ske2Grid, a new representation learning framework for improved skeleton-based action recognition. In Ske2Grid, we define a regular convolution operation upon a novel grid representation of human skeleton, which is a compact image-like grid patch constructed and learned through three novel designs. Specifically, we propose a graph-node index transform (GIT) to construct a regular grid patch through assigning the nodes in the skeleton graph one by one to the desired grid cells. To ensure that GIT is a bijection and enrich the expressiveness of the grid representation, an up-sampling transform (UPT) is learned to interpolate the skeleton graph nodes for filling the grid patch to the full. To resolve the problem when the one-step UPT is aggressive and further exploit the representation capability of the grid patch with increasing spatial size, a progressive learning strategy (PLS) is proposed which decouples the UPT into multiple steps and aligns them to multiple paired GITs through a compact cascaded design learned progressively. We construct networks upon prevailing graph convolution n

etworks and conduct experiments on six mainstream skeleton-based action recognit
ion datasets. Experiments show that our Ske2Grid significantly outperforms exist
ing GCN-based solutions under different benchmark settings, without bells and wh
istles. Code and models are available at https://github.com/OSVAI/Ske2Grid.
****

Extrapolated Random Tree for Regression
Yuchao Cai, Yuheng Ma, Yiwei Dong, Hanfang Yang
In this paper, we propose a novel tree-based algorithm named Extrapolated Random
 Tree for Regression (ERTR) that adapts to arbitrary smoothness of the regressio
n function while maintaining the interpretability of the tree. We first put forw
ard the homothetic random tree for regression (HRTR) that converges to the targe
t function as the homothetic ratio approaches zero. Then ERTR uses a linear regr
ession model to extrapolate HRTR estimations with different ratios to the ratio
zero. From the theoretical perspective, we for the first time establish the opti
mal convergence rates for ERTR when the target function resides in the general H
ölder space $C^{k,\alpha}$ for $k\in \mathbb{N}$, whereas the lower bound of the
 convergence rate of the random tree for regression (RTR) is strictly slower tha
n ERTR in the space $C^{k,\alpha}$ for $k\geq 1$. This shows that ERTR outperfor
ms RTR for the target function with high-order smoothness due to the extrapolati
on. In the experiments, we compare ERTR with state-of-the-art tree algorithms on
 real datasets to show the superior performance of our model. Moreover, promisin
g improvements are brought by using the extrapolated trees as base learners in t
he extension of ERTR to ensemble methods.
****

Cyclic Block Coordinate Descent With Variance Reduction for Composite Nonconvex
Optimization
Xufeng Cai, Chaobing Song, Stephen Wright, Jelena Diakonikolas
Nonconvex optimization is central in solving many machine learning problems, in
which block-wise structure is commonly encountered. In this work, we propose cyc
lic block coordinate methods for nonconvex optimization problems with non-asympt
otic gradient norm guarantees. Our convergence analysis is based on a gradient L
ipschitz condition with respect to a Mahalanobis norm, inspired by a recent prog
ress on cyclic block coordinate methods. In deterministic settings, our converge
nce guarantee matches the guarantee of (full-gradient) gradient descent, but wit
h the gradient Lipschitz constant being defined w.r.t. a Mahalanobis norm. In st
ochastic settings, we use recursive variance reduction to decrease the per-itera
tion cost and match the arithmetic operation complexity of current optimal stoch
astic full-gradient methods, with a unified analysis for both finite-sum and inf
inite-sum cases. We prove a faster linear convergence result when a Polyak-■ojas
iewicz (P■) condition holds. To our knowledge, this work is the first to provide
 non-asymptotic convergence guarantees — variance-reduced or not — for a cyclic
block coordinate method in general composite (smooth + nonsmooth) nonconvex sett
ings. Our experimental results demonstrate the efficacy of the proposed cyclic s
cheme in training deep neural nets.
****

Robust Weight Signatures: Gaining Robustness as Easy as Patching Weights?
Ruisi Cai, Zhenyu Zhang, Zhangyang Wang
Given a robust model trained to be resilient to one or multiple types of distrib
ution shifts (e.g., natural image corruptions), how is that "robustness" encoded
 in the model weights, and how easily can it be disentangled and/or "zero-shot"
transferred to some other models? This paper empirically suggests a surprisingly
 simple answer: linearly - by straightforward model weight arithmetic! We start
by drawing several key observations: (i) assuming that we train the same model a
rchitecture on both a clean dataset and its corrupted version, a comparison betw
een the two resultant models shows their weights to mostly differ in shallow lay
ers; (ii) the weight difference after projection, which we call "Robust Weight S
ignature" (RWS), appears to be discriminative and indicative of different corrup
tion types; (iii) perhaps most strikingly, for the same corruption type, the RWS
s obtained by one model architecture are highly consistent and transferable acro
ss different datasets. Based on those RWS observations, we propose a minimalisti

c model robustness "patching" framework that carries a model trained on clean da ta together with its pre-extracted RWSs. In this way, injecting certain robustne ss to the model is reduced to directly adding the corresponding RWS to its weigh t. We experimentally verify our proposed framework to be remarkably (1) lightwei ght. since RWSs concentrate on the shallowest few layers and we further show the y can be painlessly quantized, storing an RWS is up to 13 x more compact than st oring the full weight copy; (2) in-situ adjustable. RWSs can be appended as need ed and later taken off to restore the intact clean model. We further demonstrate one can linearly re-scale the RWS to control the patched robustness strength; (3) composable. Multiple RWSs can be added simultaneously to patch more comprehen sive robustness at once; and (4) transferable. Even when the clean model backbon e is continually adapted or updated, RWSs remain as effective patches due to the ir outstanding cross-dataset transferability.
****

Doubly Optimal No-Regret Learning in Monotone Games
Yang Cai, Weiqiang Zheng
We consider online learning in multi-player smooth monotone games. Existing algo rithms have limitations such as (1) being only applicable to strongly monotone g ames; (2) lacking the no-regret guarantee; (3) having only asymptotic or slow $\ mathcal{O}(\frac{1}{\sqrt{T}})$ last-iterate convergence rate to a Nash equilibr ium. While the $\mathcal{O}(\frac{1}{\sqrt{T}})$ rate is tight for a large class of algorithms including the well-studied extragradient algorithm and optimistic gradient algorithm, it is not optimal for all gradient-based algorithms. We pro pose the accelerated optimistic gradient (AOG) algorithm, the first doubly optim al no-regret learning algorithm for smooth monotone games. Namely, our algorithm achieves both (i) the optimal $\mathcal{O}(\sqrt{T})$ regret in the adversarial setting under smooth and convex loss functions and (ii) the optimal $\mathcal{O }(\frac{1}{T})$ last-iterate convergence rate to a Nash equilibrium in multi-pla yer smooth monotone games. As a byproduct of the accelerated last-iterate conver gence rate, we further show that each player suffers only an $\mathcal{O}(\log T)$ individual worst-case dynamic regret, providing an exponential improvement ov er the previous state-of-the-art $\mathcal{O}(\sqrt{T})$ bound.
****

Multi-Agent Learning from Learners
Mine Melodi Caliskan, Francesco Chini, Setareh Maghsudi
A large body of the "Inverse Reinforcement Learning" (IRL) literature focuses on recovering the reward function from a set of demonstrations of an expert agent who acts optimally or noisily optimally. Nevertheless, some recent works move aw ay from the optimality assumption to study the "Learning from a Learner (LfL)" p roblem, where the challenge is inferring the reward function of a learning agent from a sequence of demonstrations produced by progressively improving policies. In this work, we take one of the initial steps in addressing the multi-agent ve rsion of this problem and propose a new algorithm, MA-LfL (Multiagent Learning f rom a Learner). Unlike the state-of-the-art literature, which recovers the rewar d functions from trajectories produced by agents in some equilibrium, we study t he problem of inferring the reward functions of interacting agents in a general sum stochastic game without assuming any equilibrium state. The MA-LfL algorithm is rigorously built on a theoretical result that ensures its validity in the ca se of agents learning according to a multi-agent soft policy iteration scheme. W e empirically test MA-LfL and we observe high positive correlation between the r ecovered reward functions and the ground truth.
****

Efficient Learning of Mesh-Based Physical Simulation with Bi-Stride Multi-Scale Graph Neural Network
Yadi Cao, Menglei Chai, Minchen Li, Chenfanfu Jiang
Learning the long-range interactions on large-scale mesh-based physical systems with flat Graph Neural Networks (GNNs) and stacking Message Passings (MPs) is ch allenging due to the scaling complexity w.r.t. the number of nodes and over-smoo thing. Therefore, there has been growing interest in the community to introduce multi-scale structures to GNNs for physics simulation. However, current state-of

-the-art methods are limited by their reliance on the labor-heavy drawing of coa rser meshes or building coarser levels based on spatial proximity, which can int roduce wrong edges across geometry boundaries. Inspired by the bipartite graph d etermination, we propose a novel pooling strategy, bi-stride to tackle the afore mentioned limitations. Bi-stride pools nodes on every other frontier of the Brea dth-First-Search (BFS), without the need for the manual drawing of coarser meshe s and, avoid wrong edges introduced by spatial proximity. Additionally, it enabl es a reduced number of MP times on each level and the non-parametrized pooling a nd unpooling by interpolations, similar to convolutional Neural Networks (CNNs), which significantly reduces computational requirements. Experiments show that t he proposed framework, BSMS-GNN, significantly outperforms existing methods in t erms of both accuracy and computational efficiency in representative physics-bas ed simulation scenarios.

****

Variational Sparse Inverse Cholesky Approximation for Latent Gaussian Processes via Double Kullback-Leibler Minimization

Jian Cao, Myeongjong Kang, Felix Jimenez, Huiyan Sang, Florian Tobias Schaefer, Matthias Katzfuss

To achieve scalable and accurate inference for latent Gaussian processes, we pro pose a variational approximation based on a family of Gaussian distributions who se covariance matrices have sparse inverse Cholesky (SIC) factors. We combine th is variational approximation of the posterior with a similar and efficient SIC-r estricted Kullback-Leibler-optimal approximation of the prior. We then focus on a particular SIC ordering and nearest-neighbor-based sparsity pattern resulting in highly accurate prior and posterior approximations. For this setting, our var iational approximation can be computed via stochastic gradient descent in polylo garithmic time per iteration. We provide numerical comparisons showing that the proposed double-Kullback-Leibler-optimal Gaussian-process approximation (DKLGP) can sometimes be vastly more accurate for stationary kernels than alternative ap proaches such as inducing-point and mean-field approximations at similar computa tional complexity.

****

Learning Lightweight Object Detectors via Multi-Teacher Progressive Distillation

Shengcao Cao, Mengtian Li, James Hays, Deva Ramanan, Yu-Xiong Wang, Liangyan Gui

Resource-constrained perception systems such as edge computing and vision-for-ro botics require vision models to be both accurate and lightweight in computation and memory usage. While knowledge distillation is a proven strategy to enhance t he performance of lightweight classification models, its application to structur ed outputs like object detection and instance segmentation remains a complicated task, due to the variability in outputs and complex internal network modules in volved in the distillation process. In this paper, we propose a simple yet surpr isingly effective sequential approach to knowledge distillation that progressive ly transfers the knowledge of a set of teacher detectors to a given lightweight student. To distill knowledge from a highly accurate but complex teacher model, we construct a sequence of teachers to help the student gradually adapt. Our pro gressive strategy can be easily combined with existing detection distillation me chanisms to consistently maximize student performance in various settings. To th e best of our knowledge, we are the first to successfully distill knowledge from Transformer-based teacher detectors to convolution-based students, and unpreced entedly boost the performance of ResNet-50 based RetinaNet from 36.5% to 42.0% A P and Mask R-CNN from 38.2% to 42.5% AP on the MS COCO benchmark. Code available at https://github.com/Shengcao-Cao/MTPD.

****

One-sided Matrix Completion from Two Observations Per Row

Steven Cao, Percy Liang, Gregory Valiant

Given only a few observed entries from a low-rank matrix X, matrix completion is the problem of imputing the missing entries, and it formalizes a wide range o f real-world settings that involve estimating missing data. However, when there are too few observed entries to complete the matrix, what other aspects of the u nderlying matrix can be reliably recovered? We study one such problem setting, t

hat of “one-sided” matrix completion, where our goal is to recover the right sin gular vectors of X, even in the regime where recovering the left singular vect ors is impossible, which arises when there are more rows than columns and very f ew observations. We propose a natural algorithm that involves imputing the missi ng values of the matrix X^TX and show that even with only two observations per row in X, we can provably recover X^TX as long as we have at least $\Omega(r^2 d \log d)$ rows, where r is the rank and d is the number of columns. We evaluate our algorithm on one-sided recovery of synthetic data and low-coverage genome sequencing. In these settings, our algorithm substantially outperforms st andard matrix completion and a variety of direct factorization methods.
****

State and parameter learning with PARIS particle Gibbs
Gabriel Cardoso, Yazid Janati El Idrissi, Sylvain Le Corff, Eric Moulines, Jimmy Olsson
Non-linear state-space models, also known as general hidden Markov models (HMM), are ubiquitous in statistical machine learning, being the most classical genera tive models for serial data and sequences. Learning in HMM, either via Maximum L ikelihood Estimation (MLE) or Markov Score Climbing (MSC) requires the estimatio n of the- smoothing expectation of some additive functionals. Controlling the bi as and the variance of this estimation is crucial to establish the convergence o f learning algorithms. Our first contribution is to design a novel additive smoo thing algorithm, the Parisian particle Gibbs (PPG) sampler, which can be viewed as a PaRIS (Olsson, Westerborn 2017) algorithm driven by conditional SMC moves, resulting in bias-reduced estimates of the targeted quantities. We substantiate the PPG algorithm with theoretical results, including new bounds on bias and var iance as well as deviation inequalities. We then establish, in the learning cont ext, and under standard assumptions, non-asymptotic bounds highlighting the valu e of bias reduction and the implicit Rao–Blackwellization of PPG. These are the first non-asymptotic results of this kind in this setting. We illustrate our the oretical results with numerical experiments supporting our claims.
****

Grounding Large Language Models in Interactive Environments with Online Reinforc ement Learning
Thomas Carta, Clément Romac, Thomas Wolf, Sylvain Lamprier, Olivier Sigaud, Pier re-Yves Oudeyer
Recent works successfully leveraged Large Language Models’ (LLM) abilities to ca pture abstract knowledge about world’s physics to solve decision-making problems . Yet, the alignment between LLMs’ knowledge and the environment can be wrong an d limit functional competence due to lack of grounding. In this paper, we study an approach (named GLAM) to achieve this alignment through functional grounding: we consider an agent using an LLM as a policy that is progressively updated as the agent interacts with the environment, leveraging online Reinforcement Learni ng to improve its performance to solve goals. Using an interactive textual envir onment designed to study higher-level forms of functional grounding, and a set o f spatial and navigation tasks, we study several scientific questions: 1) Can LL Ms boost sample efficiency for online learning of various RL tasks? 2) How can i t boost different forms of generalization? 3) What is the impact of online learn ing? We study these questions by functionally grounding several variants (size, architecture) of FLAN-T5.
****

Stein Variational Goal Generation for adaptive Exploration in Multi-Goal Reinfor cement Learning
Nicolas Castanet, Olivier Sigaud, Sylvain Lamprier
In multi-goal Reinforcement Learning, an agent can share experience between rela ted training tasks, resulting in better generalization for new tasks at test tim e. However, when the goal space has discontinuities and the reward is sparse, a majority of goals are difficult to reach. In this context, a curriculum over goa ls helps agents learn by adapting training tasks to their current capabilities. In this work, we propose Stein Variational Goal Generation (SVGG), which samples goals of intermediate difficulty for the agent, by leveraging a learned predict

ive model of its goal reaching capabilities. The distribution of goals is modeled with particles that are attracted in areas of appropriate difficulty using Stein Variational Gradient Descent. We show that SVGG outperforms state-of-the-art multi-goal Reinforcement Learning methods in terms of success coverage in hard exploration problems, and demonstrate that it is endowed with a useful recovery property when the environment changes.

****

Scalable Safe Policy Improvement via Monte Carlo Tree Search

Alberto Castellini, Federico Bianchi, Edoardo Zorzi, Thiago D. Simão, Alessandro Farinelli, Matthijs T. J. Spaan

Algorithms for safely improving policies are important to deploy reinforcement learning approaches in real-world scenarios. In this work, we propose an algorithm, called MCTS-SPIBB, that computes safe policy improvement online using a Monte Carlo Tree Search based strategy. We theoretically prove that the policy generated by MCTS-SPIBB converges, as the number of simulations grows, to the optimal safely improved policy generated by Safe Policy Improvement with Baseline Bootstrapping (SPIBB), a popular algorithm based on policy iteration. Moreover, our empirical analysis performed on three standard benchmark domains shows that MCTS-SPIBB scales to significantly larger problems than SPIBB because it computes the policy online and locally, i.e., only in the states actually visited by the agent.

****

LESS-VFL: Communication-Efficient Feature Selection for Vertical Federated Learning

Timothy Castiglia, Yi Zhou, Shiqiang Wang, Swanand Kadhe, Nathalie Baracaldo, Stacy Patterson

We propose LESS-VFL, a communication-efficient feature selection method for distributed systems with vertically partitioned data. We consider a system of a server and several parties with local datasets that share a sample ID space but have different feature sets. The parties wish to collaboratively train a model for a prediction task. As part of the training, the parties wish to remove unimportant features in the system to improve generalization, efficiency, and explainability. In LESS-VFL, after a short pre-training period, the server optimizes its part of the global model to determine the relevant outputs from party models. This information is shared with the parties to then allow local feature selection without communication. We analytically prove that LESS-VFL removes spurious features from model training. We provide extensive empirical evidence that LESS-VFL can achieve high accuracy and remove spurious features at a fraction of the communication cost of other feature selection approaches.

****

On the Robustness of Text Vectorizers

Rémi Catellier, Samuel Vaiter, Damien Garreau

A fundamental issue in machine learning is the robustness of the model with respect to changes in the input. In natural language processing, models typically contain a first embedding layer, transforming a sequence of tokens into vector representations. While the robustness with respect to changes of continuous inputs is well-understood, the situation is less clear when considering discrete changes, for instance replacing a word by another in an input sentence. Our work formally proves that popular embedding schemes, such as concatenation, TF-IDF, and Paragraph Vector (a.k.a. doc2vec), exhibit robustness in the Hölder or Lipschitz sense with respect to the Hamming distance. We provide quantitative bounds for these schemes and demonstrate how the constants involved are affected by the length of the document. These findings are exemplified through a series of numerical examples.

****

Learning Globally Smooth Functions on Manifolds

Juan Cervino, Luiz F. O. Chamon, Benjamin David Haeffele, Rene Vidal, Alejandro Ribeiro

Smoothness and low dimensional structures play central roles in improving generalization and stability in learning and statistics. This work combines techniques

from semi-infinite constrained learning and manifold regularization to learn representations that are globally smooth on a manifold. To do so, it shows that under typical conditions the problem of learning a Lipschitz continuous function on a manifold is equivalent to a dynamically weighted manifold regularization problem. This observation leads to a practical algorithm based on a weighted Laplacian penalty whose weights are adapted using stochastic gradient techniques. It is shown that under mild conditions, this method estimates the Lipschitz constant of the solution, learning a globally smooth solution as a byproduct. Experiments on real world data illustrate the advantages of the proposed method relative to existing alternatives. Our code is available at https://github.com/JuanCervino/smoothbench.

****

Tighter Lower Bounds for Shuffling SGD: Random Permutations and Beyond

Jaeyoung Cha, Jaewook Lee, Chulhee Yun

We study convergence lower bounds of without-replacement stochastic gradient descent (SGD) for solving smooth (strongly-)convex finite-sum minimization problems. Unlike most existing results focusing on final iterate lower bounds in terms of the number of components n and the number of epochs K, we seek bounds for arbitrary weighted average iterates that are tight in all factors including the condition number κ. For SGD with Random Reshuffling, we present lower bounds that have tighter κ dependencies than existing bounds. Our results are the first to perfectly close the gap between lower and upper bounds for weighted average iterates in both strongly-convex and convex cases. We also prove weighted average iterate lower bounds for arbitrary permutation-based SGD, which apply to all variants that carefully choose the best permutation. Our bounds improve the existing bounds in factors of n and κ and thereby match the upper bounds shown for a recently proposed algorithm called GraB.

****

Orthogonality-Enforced Latent Space in Autoencoders: An Approach to Learning Disentangled Representations

Jaehoon Cha, Jeyan Thiyagalingam

Noting the importance of factorizing (or disentangling) the latent space, we propose a novel, non-probabilistic disentangling framework for autoencoders, based on the principles of symmetry transformations that are independent of one another. To the best of our knowledge, this is the first deterministic model that is aiming to achieve disentanglement based on autoencoders using only a reconstruction loss without pairs of images or labels, by explicitly introducing inductive biases into a model architecture through Euler encoding. The proposed model is then compared with a number of state-of-the-art models, relevant to disentanglement, including symmetry-based models and generative models. Our evaluation using six different disentanglement metrics, including the unsupervised disentanglement metric we propose here in this paper, shows that the proposed model can offer better disentanglement, especially when variances of the features are different, where other methods may struggle. We believe that this model opens several opportunities for linear disentangled representation learning based on deterministic autoencoders.

****

STEERING : Stein Information Directed Exploration for Model-Based Reinforcement Learning

Souradip Chakraborty, Amrit Bedi, Alec Koppel, Mengdi Wang, Furong Huang, Dinesh Manocha

Directed Exploration is a crucial challenge in reinforcement learning (RL), especially when rewards are sparse. Information-directed sampling (IDS), which optimizes the information ratio, seeks to do so by augmenting regret with information gain. However, estimating information gain is computationally intractable or relies on restrictive assumptions which prohibit its use in many practical instances. In this work, we posit an alternative exploration incentive in terms of the integral probability metric (IPM) between a current estimate of the transition model and the unknown optimal, which under suitable conditions, can be computed in closed form with the kernelized Stein discrepancy (KSD). Based on KSD, we deve

lop a novel algorithm STEERING: STEin information dirEcted exploration for model-based Reinforcement LearnING. To enable its derivation, we develop fundamentally new variants of KSD for discrete conditional distributions. We further establish that STEERING archives sublinear Bayesian regret, improving upon prior learning rates of information-augmented MBRL, IDS included. Experimentally, we show that the proposed algorithm is computationally affordable and outperforms several prior approaches.

****

Thompson Sampling for High-Dimensional Sparse Linear Contextual Bandits

Sunrit Chakraborty, Saptarshi Roy, Ambuj Tewari

We consider the stochastic linear contextual bandit problem with high-dimensional features. We analyze the Thompson sampling algorithm using special classes of sparsity-inducing priors (e.g., spike-and-slab) to model the unknown parameter and provide a nearly optimal upper bound on the expected cumulative regret. To the best of our knowledge, this is the first work that provides theoretical guarantees of Thompson sampling in high-dimensional and sparse contextual bandits. For faster computation, we use variational inference instead of Markov Chain Monte Carlo (MCMC) to approximate the posterior distribution. Extensive simulations demonstrate the improved performance of our proposed algorithm over existing ones.

****

Representations and Exploration for Deep Reinforcement Learning using Singular Value Decomposition

Yash Chandak, Shantanu Thakoor, Zhaohan Daniel Guo, Yunhao Tang, Remi Munos, Will Dabney, Diana L Borsa

Representation learning and exploration are among the key challenges for any deep reinforcement learning agent. In this work, we provide a singular value decomposition based method that can be used to obtain representations that preserve the underlying transition structure in the domain. Perhaps interestingly, we show that these representations also capture the relative frequency of state visitations, thereby providing an estimate for pseudo-counts for free. To scale this decomposition method to large-scale domains, we provide an algorithm that never requires building the transition matrix, can make use of deep networks, and also permits mini-batch training. Further, we draw inspiration from predictive state representations and extend our decomposition method to partially observable environments. With experiments on multi-task settings with partially observable domains, we show that the proposed method can not only learn useful representation on DM-Lab-30 environments (that have inputs involving language instructions, pixel images, rewards, among others) but it can also be effective at hard exploration tasks in DM-Hard-8 environments.

****

Memory-Based Dual Gaussian Processes for Sequential Learning

Paul Edmund Chang, Prakhar Verma, S. T. John, Arno Solin, Mohammad Emtiyaz Khan

Sequential learning with Gaussian processes (GPs) is challenging when access to past data is limited, for example, in continual and active learning. In such cases, errors can accumulate over time due to inaccuracies in the posterior, hyperparameters, and inducing points, making accurate learning challenging. Here, we present a method to keep all such errors in check using the recently proposed dual sparse variational GP. Our method enables accurate inference for generic likelihoods and improves learning by actively building and updating a memory of past data. We demonstrate its effectiveness in several applications involving Bayesian optimization, active learning, and continual learning.

****

Muse: Text-To-Image Generation via Masked Generative Transformers

Huiwen Chang, Han Zhang, Jarred Barber, Aaron Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Patrick Murphy, William T. Freeman, Michael Rubinstein, Yuanzhen Li, Dilip Krishnan

We present Muse, a text-to-image Transformermodel that achieves state-of-the-art image genera-tion performance while being significantly moreefficient than diffusion or autoregressive models.Muse is trained on a masked modeling task indiscrete token space: given the text embeddingextracted from a pre-trained large lang

uage model(LLM), Muse learns to predict randomly maskedimage tokens. Compared to pixel-space diffusionmodels, such as Imagen and DALL-E 2, Muse issignificantly more efficient due to the use of dis-crete tokens and requires fewer sampling it era-tions; compared to autoregressive models such asParti, Muse is more efficien t due to the use of par-allel decoding. The use of a pre-trained LLM en-ables fi ne-grained language understanding, whichtranslates to high-fidelity image genera tion andthe understanding of visual concepts such as ob-jects, their spatial rel ationships, pose, cardinalityetc. Our 900M parameter model achieves a newSOTA on CC3M, with an FID score of 6.06. TheMuse 3B parameter model achieves an FID of7 .88 on zero-shot COCO evaluation, along with aCLIP score of 0.32. Muse also dire ctly enables anumber of image editing applications without theneed to fine-tune or invert the model: inpainting,outpainting, and mask-free editing. More results and videos demonstrating editing are available at https://muse-icml.github.io/
****

On Investigating the Conservative Property of Score-Based Generative Models
Chen-Hao Chao, Wei-Fang Sun, Bo-Wun Cheng, Chun-Yi Lee
Existing Score-Based Models (SBMs) can be categorized into constrained SBMs (CSB Ms) or unconstrained SBMs (USBMs) according to their parameterization approaches . CSBMs model probability density functions as Boltzmann distributions, and assi gn their predictions as the negative gradients of some scalar-valued energy func tions. On the other hand, USBMs employ flexible architectures capable of directl y estimating scores without the need to explicitly model energy functions. In th is paper, we demonstrate that the architectural constraints of CSBMs may limit t heir modeling ability. In addition, we show that USBMs' inability to preserve th e property of conservativeness may lead to degraded performance in practice. To address the above issues, we propose Quasi-Conservative Score-Based Models (QCSB Ms) for keeping the advantages of both CSBMs and USBMs. Our theoretical derivati ons demonstrate that the training objective of QCSBMs can be efficiently integra ted into the training processes by leveraging the Hutchinson's trace estimator. In addition, our experimental results on the CIFAR-10, CIFAR-100, ImageNet, and SVHN datasets validate the effectiveness of QCSBMs. Finally, we justify the adva ntage of QCSBMs using an example of a one-layered autoencoder.
****

Robust and private stochastic linear bandits
Vasileios Charisopoulos, Hossein Esfandiari, Vahab Mirrokni
In this paper, we study the stochastic linear bandit problem under the additiona l requirements of differential privacy, robustness and batched observations. In particular, we assume an adversary randomly chooses a constant fraction of the o bserved rewards in each batch, replacing them with arbitrary numbers. We present differentially private and robust variants of the arm elimination algorithm usi ng logarithmic batch queries under two privacy models and provide regret bounds in both settings. In the first model, every reward in each round is reported by a potentially different client, which reduces to standard local differential pri vacy (LDP). In the second model, every action is "owned" by a different client, who may aggregate the rewards over multiple queries and privatize the aggregate response instead. To the best of our knowledge, our algorithms are the first sim ultaneously providing differential privacy and adversarial robustness in the sto chastic linear bandits problem.
****

Streaming Submodular Maximization with Differential Privacy
Anamay Chaturvedi, Huy Nguyen, Thy Dinh Nguyen
In this work, we study the problem of privately maximizing a submodular function in the streaming setting. Extensive work has been done on privately maximizing submodular functions in the general case when the function depends upon the priv ate data of individuals. However, when the size of the data stream drawn from th e domain of the objective function is large or arrives very fast, one must priva tely optimize the objective within the constraints of the streaming setting. We establish fundamental differentially private baselines for this problem and then derive better trade-offs between privacy and utility for the special case of de composable submodular functions. A submodular function is decomposable when it c

an be written as a sum of submodular functions; this structure arises naturally when each summand function models the utility of an individual and the goal is t o study the total utility of the whole population as in the well-known Combinato rial Public Projects Problem. Finally, we complement our theoretical analysis wi th experimental corroboration.
****

Why does Throwing Away Data Improve Worst-Group Error?
Kamalika Chaudhuri, Kartik Ahuja, Martin Arjovsky, David Lopez-Paz
When facing data with imbalanced classes or groups, practitioners follow an intr iguing strategy to achieve best results. They throw away examples until the clas ses or groups are balanced in size, and then perform empirical risk minimization on the reduced training set. This opposes common wisdom in learning theory, whe re the expected error is supposed to decrease as the dataset grows in size. In t his work, we leverage extreme value theory to address this apparent contradictio n. Our results show that the tails of the data distribution play an important ro le in determining the worst-group-accuracy of linear classifiers. When learning on data with heavy tails, throwing away data restores the geometric symmetry of the resulting classifier, and therefore improves its worst-group generalization.
****

Collaborative Multi-Agent Heterogeneous Multi-Armed Bandits
Ronshee Chawla, Daniel Vial, Sanjay Shakkottai, R. Srikant
The study of collaborative multi-agent bandits has attracted significant attenti on recently. In light of this, we initiate the study of a new collaborative sett ing, consisting of N agents such that each agent is learning one of M stocha stic multi-armed bandits to minimize their group cumulative regret. We develop d ecentralized algorithms which facilitate collaboration between the agents under two scenarios. We characterize the performance of these algorithms by deriving t he per agent cumulative regret and group regret upper bounds. We also prove lowe r bounds for the group regret in this setting, which demonstrates the near-optim al behavior of the proposed algorithms.
****

Correcting discount-factor mismatch in on-policy policy gradient methods
Fengdi Che, Gautham Vasan, A. Rupam Mahmood
The policy gradient theorem gives a convenient form of the policy gradient in te rms of three factors: an action value, a gradient of the action likelihood, and a state distribution involving discounting called the discounted stationary dist ribution. But commonly used on-policy methods based on the policy gradient theor em ignores the discount factor in the state distribution, which is technically i ncorrect and may even cause degenerate learning behavior in some environments. A n existing solution corrects this discrepancy by using γ^t as a factor in the gradient estimate. However, this solution is not widely adopted and does no t work well in tasks where the later states are similar to earlier states. We in troduce a novel distribution correction to account for the discounted stationary distribution that can be plugged into many existing gradient estimators. Our co rrection circumvents the performance degradation associated with the γ^t correction with a lower variance. Importantly, compared to the uncorrected estim ators, our algorithm provides improved state emphasis to evade suboptimal polici es in certain environments and consistently matches or exceeds the original perf ormance on several OpenAI gym and DeepMind suite benchmarks.
****

Fast Federated Machine Unlearning with Nonlinear Functional Theory
Tianshi Che, Yang Zhou, Zijie Zhang, Lingjuan Lyu, Ji Liu, Da Yan, Dejing Dou, J un Huan
Federated machine unlearning (FMU) aims to remove the influence of a specified s ubset of training data upon request from a trained federated learning model. Des pite achieving remarkable performance, existing FMU techniques suffer from ineff iciency due to two sequential operations of training and retraining/unlearning o n large-scale datasets. Our prior study, PCMU, was proposed to improve the effic iency of centralized machine unlearning (CMU) with certified guarantees, by simu ltaneously executing the training and unlearning operations. This paper proposes

a fast FMU algorithm, FFMU, for improving the FMU efficiency while maintaining the unlearning quality. The PCMU method is leveraged to train a local machine learning (MU) model on each edge device. We propose to employ nonlinear functional analysis techniques to refine the local MU models as output functions of a Nemytskii operator. We conduct theoretical analysis to derive that the Nemytskii operator has a global Lipschitz constant, which allows us to bound the difference between two MU models regarding the distance between their gradients. Based on the Nemytskii operator and average smooth local gradients, the global MU model on the server is guaranteed to achieve close performance to each local MU model with the certified guarantees.

****

On the Statistical Benefits of Temporal Difference Learning

David Cheikhi, Daniel Russo

Given a dataset on actions and resulting long-term rewards, a direct estimation approach fits value functions that minimize prediction error on the training data. Temporal difference learning (TD) methods instead fit value functions by minimizing the degree of temporal inconsistency between estimates made at successive time-steps. Focusing on finite state Markov chains, we provide a crisp asymptotic theory of the statistical advantages of this approach. First, we show that an intuitive inverse trajectory pooling coefficient completely characterizes the percent reduction in mean-squared error of value estimates. Depending on problem structure, the reduction could be enormous or nonexistent. Next, we prove that there can be dramatic improvements in estimates of the difference in value-to-go for two states: TD’s errors are bounded in terms of a novel measure – the problem’s trajectory crossing time – which can be much smaller than the problem’s time horizon.

****

Multi-Layer Neural Networks as Trainable Ladders of Hilbert Spaces

Zhengdao Chen

To characterize the functions spaces explored by multi-layer neural networks (NNs), we introduce Neural Hilbert Ladders (NHLs), a collection of reproducing kernel Hilbert spaces (RKHSes) that are defined iteratively and adaptive to training. First, we prove a correspondence between functions expressed by L-layer NNs and those belonging to L-level NHLs. Second, we prove generalization guarantees for learning the NHL based on a new complexity measure. Third, corresponding to the training of multi-layer NNs in the infinite-width mean-field limit, we derive an evolution of the NHL characterized by the dynamics of multiple random fields. Finally, we examine linear and shallow NNs from the new perspective and complement the theory with numerical results.

****

Beyond the Edge of Stability via Two-step Gradient Updates

Lei Chen, Joan Bruna

Gradient Descent (GD) is a powerful workhorse of modern machine learning thanks to its scalability and efficiency in high-dimensional spaces. Its ability to find local minimisers is only guaranteed for losses with Lipschitz gradients, where it can be seen as a ’bona-fide’ discretisation of an underlying gradient flow. Yet, many ML setups involving overparametrised models do not fall into this problem class, which has motivated research beyond the so-called ”Edge of Stability” (EoS), where the step-size crosses the admissibility threshold inversely proportional to the Lipschitz constant above. Perhaps surprisingly, GD has been empirically observed to still converge regardless of local instability and oscillatory behavior. The incipient theoretical analysis of this phenomena has mainly focused in the overparametrised regime, where the effect of choosing a large learning rate may be associated to a ‘Sharpness-Minimisation’ implicit regularisation within the manifold of minimisers, under appropriate asymptotic limits. In contrast, in this work we directly examine the conditions for such unstable convergence, focusing on simple, yet representative, learning problems, via analysis of two-step gradient updates. Specifically, we characterize a local condition involving third-order derivatives that guarantees existence and convergence to fixed points of the two-step updates, and leverage such property in a teacher-student set

ting, under population loss. Finally, starting from Matrix Factorization, we pro vide observations of period-2 orbit of GD in high-dimensional settings with intu ition of its dynamics, along with exploration into more general settings.
****

Trompt: Towards a Better Deep Neural Network for Tabular Data
Kuan-Yu Chen, Ping-Han Chiang, Hsin-Rung Chou, Ting-Wei Chen, Tien-Hao Chang
Tabular data is arguably one of the most commonly used data structures in variou s practical domains, including finance, healthcare and e-commerce. The inherent heterogeneity allows tabular data to store rich information. However, based on a recently published tabular benchmark, we can see deep neural networks still fal l behind tree-based models on tabular datasets. In this paper, we propose Trompt –which stands for Tabular Prompt–a novel architecture inspired by prompt learnin g of language models. The essence of prompt learning is to adjust a large pre-tr ained model through a set of prompts outside the model without directly modifyin g the model. Based on this idea, Trompt separates the learning strategy of tabul ar data into two parts. The first part, analogous to pre-trained models, focus o n learning the intrinsic information of a table. The second part, analogous to p rompts, focus on learning the variations among samples. Trompt is evaluated with the benchmark mentioned above. The experimental results demonstrate that Trompt outperforms state-of-the-art deep neural networks and is comparable to tree-bas ed models.
****

Differentially Private Stochastic Convex Optimization under a Quantile Loss Func tion
Du Chen, Geoffrey A. Chua
We study (ε,δ)-differentially private (DP) stochastic convex op timization under an r-th quantile loss function taking the form $c(u) = ru^+ + (1-r)(-u)^+$. The function is non-smooth, and we propose to approximate it with a smooth function obtained by convolution smoothing, which enjoys both structur e and bandwidth flexibility and can address outliers. This leads to a better app roximation than those obtained from existing methods such as Moreau Envelope. We then design private algorithms based on DP stochastic gradient descent and obje ctive perturbation, and show that both algorithms achieve (near) optimal excess generalization risk $O(\max\{\frac{1}{\sqrt{n}}, \frac{\sqrt{d\ln(1/\delta)}}{n\ varepsilon}\})$. Through objective perturbation, we further derive an upper boun d $O(\max\{\sqrt{\frac{d}{n}}, \sqrt{\frac{d\ln(1/\delta)}{n\varepsilon}}\})$ on the parameter estimation error under mild assumptions on data generating proces ses. Some applications in private quantile regression and private inventory cont rol will be discussed.
****

Restoration-Degradation Beyond Linear Diffusions: A Non-Asymptotic Analysis For DDIM-type Samplers
Sitan Chen, Giannis Daras, Alex Dimakis
We develop a framework for non-asymptotic analysis of deterministic samplers use d for diffusion generative modeling. Several recent works have analyzed stochast ic samplers using tools like Girsanov’s theorem and a chain rule variant of the interpolation argument. Unfortunately, these techniques give vacuous bounds when applied to deterministic samplers. We give a new operational interpretation for deterministic sampling by showing that one step along the probability flow ODE can be expressed as two steps: 1) a restoration step that runs gradient ascent o n the conditional log-likelihood at some infinitesimally previous time, and 2) a degradation step that runs the forward process using noise pointing back toward s the current iterate. This perspective allows us to extend denoising diffusion implicit models to general, non-linear forward processes. We then develop the fi rst polynomial convergence bounds for these samplers under mild conditions on th e data distribution.
****

Provably Convergent Schrödinger Bridge with Applications to Probabilistic Time S eries Imputation
Yu Chen, Wei Deng, Shikai Fang, Fengpei Li, Nicole Tianjiao Yang, Yikai Zhang, K

ashif Rasul, Shandian Zhe, Anderson Schneider, Yuriy Nevmyvaka
The Schrödinger bridge problem (SBP) is gaining increasing attention in generati ve modeling and showing promising potential even in comparison with the score-ba sed generative models (SGMs). SBP can be interpreted as an entropy-regularized o ptimal transport problem, which conducts projections onto every other marginal a lternatingly. However, in practice, only approximated projections are accessible and their convergence is not well understood. To fill this gap, we present a fi rst convergence analysis of the Schrödinger bridge algorithm based on approximat ed projections. As for its practical applications, we apply SBP to probabilistic time series imputation by generating missing values conditioned on observed dat a. We show that optimizing the transport cost improves the performance and the p roposed algorithm achieves the state-of-the-art result in healthcare and environ mental data while exhibiting the advantage of exploring both temporal and featur e patterns in probabilistic time series imputation.
****
ED-Batch: Efficient Automatic Batching of Dynamic Neural Networks via Learned Fi nite State Machines
Siyuan Chen, Pratik Pramod Fegade, Tianqi Chen, Phillip Gibbons, Todd Mowry
Batching has a fundamental influence on the efficiency of deep neural network (D NN) execution. However, for dynamic DNNs, efficient batching is particularly cha llenging as the dataflow graph varies per input instance. As a result, state-of- the-art frameworks use heuristics that result in suboptimal batching decisions. Further, batching puts strict restrictions on memory adjacency and can lead to h igh data movement costs. In this paper, we provide an approach for batching dyna mic DNNs based on finite state machines, which enables the automatic discovery o f batching policies specialized for each DNN via reinforcement learning. Moreove r, we find that memory planning that is aware of the batching policy can save si gnificant data movement overheads, which is automated by a PQ tree-based algorit hm we introduce. Experimental results show that our framework speeds up state-of -the-art frameworks by on average 1.15x, 1.39x, and 2.45x for chain-based, tree- based, and lattice-based DNNs across CPU and GPU. The framework is open-sourced at https://github.com/gulang2019/ED-Batch.git.
****
Is Learning Summary Statistics Necessary for Likelihood-free Inference?
Yanzhi Chen, Michael U. Gutmann, Adrian Weller
Likelihood-free inference (LFI) is a set of techniques for inference in implicit statistical models. A longstanding question in LFI has been how to design or le arn good summary statistics of data, but this might now seem unnecessary due to the advent of recent end-to-end (i.e. neural network-based) LFI methods. In this work, we rethink this question with a new method for learning summary statistic s. We show that learning sufficient statistics may be easier than direct posteri or inference, as the former problem can be reduced to a set of low-dimensional, easy-to-solve learning problems. This suggests us to explicitly decouple summary statistics learning from posterior inference in LFI. Experiments on diverse inf erence tasks with different data types validate our hypothesis.
****
Subequivariant Graph Reinforcement Learning in 3D Environments
Runfa Chen, Jiaqi Han, Fuchun Sun, Wenbing Huang
Learning a shared policy that guides the locomotion of different agents is of co re interest in Reinforcement Learning (RL), which leads to the study of morpholo gy-agnostic RL. However, existing benchmarks are highly restrictive in the choic e of starting point and target point, constraining the movement of the agents wi thin 2D space. In this work, we propose a novel setup for morphology-agnostic RL , dubbed Subequivariant Graph RL in 3D environments (3D-SGRL). Specifically, we first introduce a new set of more practical yet challenging benchmarks in 3D spa ce that allows the agent to have full Degree-of-Freedoms to explore in arbitrary directions starting from arbitrary configurations. Moreover, to optimize the po licy over the enlarged state-action space, we propose to inject geometric symmet ry, i.e., subequivariance, into the modeling of the policy and Q-function such t hat the policy can generalize to all directions, improving exploration efficienc

y. This goal is achieved by a novel SubEquivariant Transformer (SET) that permit s expressive message exchange. Finally, we evaluate the proposed method on the p roposed benchmarks, where our method consistently and significantly outperforms existing approaches on single-task, multi-task, and zero-shot generalization sce narios. Extensive ablations are also conducted to verify our design.
****

GuardHFL: Privacy Guardian for Heterogeneous Federated Learning
Hanxiao Chen, Meng Hao, Hongwei Li, Kangjie Chen, Guowen Xu, Tianwei Zhang, Xili n Zhang
Heterogeneous federated learning (HFL) enables clients with different computatio n and communication capabilities to collaboratively train their own customized m odels via a query-response paradigm on auxiliary datasets. However, such a parad igm raises serious privacy concerns due to the leakage of highly sensitive query samples and response predictions. We put forth GuardHFL, the first-of-its-kind efficient and privacy-preserving HFL framework. GuardHFL is equipped with a nove l HFL-friendly secure querying scheme built on lightweight secret sharing and sy mmetric-key techniques. The core of GuardHFL is two customized multiplication an d comparison protocols, which substantially boost the execution efficiency. Exte nsive evaluations demonstrate that GuardHFL significantly outperforms the altern ative instantiations based on existing state-of-the-art techniques in both runti me and communication cost.
****

Efficient and Degree-Guided Graph Generation via Discrete Diffusion Modeling
Xiaohui Chen, Jiaxing He, Xu Han, Liping Liu
Diffusion-based generative graph models have been proven effective in generating high-quality small graphs. However, they need to be more scalable for generatin g large graphs containing thousands of nodes desiring graph statistics. In this work, we propose EDGE, a new diffusion-based generative graph model that address es generative tasks with large graphs. To improve computation efficiency, we enc ourage graph sparsity by using a discrete diffusion process that randomly remove s edges at each time step and finally obtains an empty graph. EDGE only focuses on a portion of nodes in the graph at each denoising step. It makes much fewer e dge predictions than previous diffusion-based models. Moreover, EDGE admits expl icitly modeling the node degrees of the graphs, further improving the model perf ormance. The empirical study shows that EDGE is much more efficient than competi ng methods and can generate large graphs with thousands of nodes. It also outper forms baseline models in generation quality: graphs generated by our approach ha ve more similar graph statistics to those of the training graphs.
****

Evolving Semantic Prototype Improves Generative Zero-Shot Learning
Shiming Chen, Wenjin Hou, Ziming Hong, Xiaohan Ding, Yibing Song, Xinge You, Ton gliang Liu, Kun Zhang
In zero-shot learning (ZSL), generative methods synthesize class-related sample features based on predefined semantic prototypes. They advance the ZSL performan ce by synthesizing unseen class sample features for better training the classifi er. We observe that each class’s predefined semantic prototype (also referred to as semantic embedding or condition) does not accurately match its real semantic prototype. So the synthesized visual sample features do not faithfully represen t the real sample features, limiting the classifier training and existing ZSL pe rformance. In this paper, we formulate this mismatch phenomenon as the visual-se mantic domain shift problem. We propose a dynamic semantic prototype evolving (D SP) method to align the empirically predefined semantic prototypes and the real prototypes for class-related feature synthesis. The alignment is learned by refi ning sample features and semantic prototypes in a unified framework and making t he synthesized visual sample features approach real sample features. After align ment, synthesized sample features from unseen classes are closer to the real sam ple features and benefit DSP to improve existing generative ZSL methods by 8.5%, 8.0%, and 9.7% on the standard CUB, SUN AWA2 datasets, the significant performa nce improvement indicates that evolving semantic prototype explores a virgin fie ld in ZSL.

****
Explore and Exploit the Diverse Knowledge in Model Zoo for Domain Generalization
Yimeng Chen, Tianyang Hu, Fengwei Zhou, Zhenguo Li, Zhi-Ming Ma
The proliferation of pretrained models, as a result of advancements in pretraining techniques, has led to the emergence of a vast zoo of publicly available models. Effectively utilizing these resources to obtain models with robust out-of-distribution generalization capabilities for downstream tasks has become a crucial area of research. Previous research has primarily focused on identifying the most powerful models within the model zoo, neglecting to fully leverage the diverse inductive biases contained within. This paper argues that the knowledge contained in weaker models is valuable and presents a method for leveraging the diversity within the model zoo to improve out-of-distribution generalization capabilities. Specifically, we investigate the behaviors of various pretrained models across different domains of downstream tasks by characterizing the variations in their encoded representations in terms of two dimensions: diversity shift and correlation shift. This characterization enables us to propose a new algorithm for integrating diverse pretrained models, not limited to the strongest models, in order to achieve enhanced out-of-distribution generalization performance. Our proposed method demonstrates state-of-the-art empirical results on a variety of data sets, thus validating the benefits of utilizing diverse knowledge.
****
Decentralized Stochastic Bilevel Optimization with Improved per-Iteration Complexity
Xuxing Chen, Minhui Huang, Shiqian Ma, Krishna Balasubramanian
Bilevel optimization recently has received tremendous attention due to its great success in solving important machine learning problems like meta learning, reinforcement learning, and hyperparameter optimization. Extending single-agent training on bilevel problems to the decentralized setting is a natural generalization, and there has been a flurry of work studying decentralized bilevel optimization algorithms. However, it remains unknown how to design the distributed algorithm with sample complexity and convergence rate comparable to SGD for stochastic optimization, and at the same time without directly computing the exact Hessian or Jacobian matrices. In this paper we propose such an algorithm. More specifically, we propose a novel decentralized stochastic bilevel optimization (DSBO) algorithm that only requires first order stochastic oracle, Hessian-vector product and Jacobian-vector product oracle. The sample complexity of our algorithm matches the currently best known results for DSBO, while our algorithm does not require estimating the full Hessian and Jacobian matrices, thereby possessing to improved per-iteration complexity.
****
Score Approximation, Estimation and Distribution Recovery of Diffusion Models on Low-Dimensional Data
Minshuo Chen, Kaixuan Huang, Tuo Zhao, Mengdi Wang
Diffusion models achieve state-of-the-art performance in various generation tasks. However, their theoretical foundations fall far behind. This paper studies score approximation, estimation, and distribution recovery of diffusion models, when data are supported on an unknown low-dimensional linear subspace. Our result provides sample complexity bounds for distribution estimation using diffusion models. We show that with a properly chosen neural network architecture, the score function can be both accurately approximated and efficiently estimated. Further, the generated distribution based on the estimated score function captures the data geometric structures and converges to a close vicinity of the data distribution. The convergence rate depends on subspace dimension, implying that diffusion models can circumvent the curse of data ambient dimensionality.
****
Sample Complexity of Probability Divergences under Group Symmetry
Ziyu Chen, Markos Katsoulakis, Luc Rey-Bellet, Wei Zhu
We rigorously quantify the improvement in the sample complexity of variational divergence estimations for group-invariant distributions. In the cases of the Wasserstein-1 metric and the Lipschitz-regularized α-divergences, the reduct

ion of sample complexity is proportional to an ambient-dimension-dependent power of the group size. For the maximum mean discrepancy (MMD), the improvement of sample complexity is more nuanced, as it depends on not only the group size but also the choice of kernel. Numerical simulations verify our theories.
****
Improved Analysis of Score-based Generative Modeling: User-Friendly Bounds under Minimal Smoothness Assumptions
Hongrui Chen, Holden Lee, Jianfeng Lu
We give an improved theoretical analysis of score-based generative modeling. Under a score estimate with small L^2 error (averaged across timesteps), we provide efficient convergence guarantees for any data distribution with second-order moment, by either employing early stopping or assuming smoothness condition on the score function of the data distribution. Our result does not rely on any log-concavity or functional inequality assumption and has a logarithmic dependence on the smoothness. In particular, we show that under only a finite second moment condition, approximating the following in reverse KL divergence in ϵ-accuracy can be done in $\tilde O\left(\frac{d \log (1/\delta)}{\epsilon}\right)$ steps: 1) the variance-δ Gaussian perturbation of any data distribution; 2) data distributions with $1/\delta$-smooth score functions. Our analysis also provides a quantitative comparison between different discrete approximations and may guide the choice of discretization points in practice.
****
Bidirectional Looking with A Novel Double Exponential Moving Average to Adaptive and Non-adaptive Momentum Optimizers
Yineng Chen, Zuchao Li, Lefei Zhang, Bo Du, Hai Zhao
Optimizer is an essential component for the success of deep learning, which guides the neural network to update the parameters according to the loss on the training set. SGD and Adam are two classical and effective optimizers on which researchers have proposed many variants, such as SGDM and RAdam. In this paper, we innovatively combine the backward-looking and forward-looking aspects of the optimizer algorithm and propose a novel Admeta (A Double exponential Moving averagE To Adaptive and non-adaptive momentum) optimizer framework. For backward-looking part, we propose a DEMA variant scheme, which is motivated by a metric in the stock market, to replace the common exponential moving average scheme. While in the forward-looking part, we present a dynamic lookahead strategy which asymptotically approaches a set value, maintaining its speed at early stage and high convergence performance at final stage. Based on this idea, we provide two optimizer implementations, AdmetaR and AdmetaS, the former based on RAdam and the latter based on SGDM. Through extensive experiments on diverse tasks, we find that the proposed Admeta optimizer outperforms our base optimizers and shows advantages over recently proposed competitive optimizers. We also provide theoretical proof of these two algorithms, which verifies the convergence of our proposed Admeta.
****
HarsanyiNet: Computing Accurate Shapley Values in a Single Forward Propagation
Lu Chen, Siyu Lou, Keyan Zhang, Jin Huang, Quanshi Zhang
The Shapley value is widely regarded as a trustworthy attribution metric. However, when people use Shapley values to explain the attribution of input variables of a deep neural network (DNN), it usually requires a very high computational cost to approximate relatively accurate Shapley values in real-world applications. Therefore, we propose a novel network architecture, the HarsanyiNet, which makes inferences on the input sample and simultaneously computes the exact Shapley values of the input variables in a single forward propagation. The HarsanyiNet is designed on the theoretical foundation that the Shapley value can be reformulated as the redistribution of Harsanyi interactions encoded by the network.
****
Generalized Implicit Follow-The-Regularized-Leader
Keyi Chen, Francesco Orabona
We propose a new class of online learning algorithms, generalized implicit Follow-The-Regularized-Leader (FTRL), that expands the scope of FTRL framework. Generalized implicit FTRL can recover known algorithms, such as FTRL with linearized

losses and implicit FTRL, and it allows the design of new update rules, as extensions of aProx and Mirror-Prox to FTRL. Our theory is constructive in the sense that it provides a simple unifying framework to design updates that directly improve the worst-case upper bound on the regret. The key idea is substituting the linearization of the losses with a Fenchel-Young inequality. We show the flexibility of the framework by proving that some known algorithms, like the Mirror-Prox updates, are instantiations of the generalized implicit FTRL. Finally, the new framework allows us to recover the temporal variation bound of implicit OMD, with the same computational complexity.

****

Fisher Information Embedding for Node and Graph Learning

Dexiong Chen, Paolo Pellizzoni, Karsten Borgwardt

Attention-based graph neural networks (GNNs), such as graph attention networks (GATs), have become popular neural architectures for processing graph-structured data and learning node embeddings. Despite their empirical success, these models rely on labeled data and the theoretical properties of these models have yet to be fully understood. In this work, we propose a novel attention-based node embedding framework for graphs. Our framework builds upon a hierarchical kernel for multisets of subgraphs around nodes (e.g. neighborhoods) and each kernel leverages the geometry of a smooth statistical manifold to compare pairs of multisets, by “projecting” the multisets onto the manifold. By explicitly computing node embeddings with a manifold of Gaussian mixtures, our method leads to a new attention mechanism for neighborhood aggregation. We provide theoretical insights into generalizability and expressivity of our embeddings, contributing to a deeper understanding of attention-based GNNs. We propose both efficient unsupervised and supervised methods for learning the embeddings. Through experiments on several node classification benchmarks, we demonstrate that our proposed method outperforms existing attention-based graph models like GATs. Our code is available at https://github.com/BorgwardtLab/fisher_information_embedding.

****

Rethinking Visual Reconstruction: Experience-Based Content Completion Guided by Visual Cues

Jiaxuan Chen, Yu Qi, Gang Pan

Decoding seen images from brain activities has been an absorbing field. However, the reconstructed images still suffer from low quality with existing studies. This can be because our visual system is not like a camera that ”remembers” every pixel. Instead, only part of the information can be perceived with our selective attention, and the brain ”guesses” the rest to form what we think we see. Most existing approaches ignored the brain completion mechanism. In this work, we propose to reconstruct seen images with both the visual perception and the brain completion process, and design a simple, yet effective visual decoding framework to achieve this goal. Specifically, we first construct a shared discrete representation space for both brain signals and images. Then, a novel self-supervised token-to-token inpainting network is designed to implement visual content completion by building context and prior knowledge about the visual objects from the discrete latent space. Our approach improved the quality of visual reconstruction significantly and achieved state-of-the-art.

****

Stratified Adversarial Robustness with Rejection

Jiefeng Chen, Jayaram Raghuram, Jihye Choi, Xi Wu, Yingyu Liang, Somesh Jha

Recently, there is an emerging interest in adversarially training a classifier with a rejection option (also known as a selective classifier) for boosting adversarial robustness. While rejection can incur a cost in many applications, existing studies typically associate zero cost with rejecting perturbed inputs, which can result in the rejection of numerous slightly-perturbed inputs that could be correctly classified. In this work, we study adversarially-robust classification with rejection in the stratified rejection setting, where the rejection cost is modeled by rejection loss functions monotonically non-increasing in the perturbation magnitude. We theoretically analyze the stratified rejection setting and propose a novel defense method – Adversarial Training with Consistent Prediction-

based Rejection (CPR) – for building a robust selective classifier. Experiments on image datasets demonstrate that the proposed method significantly outperforms existing methods under strong adaptive attacks. For instance, on CIFAR-10, CPR reduces the total robust loss (for different rejection losses) by at least 7.3% under both seen and unseen attacks.

****

Multi-task Hierarchical Adversarial Inverse Reinforcement Learning

Jiayu Chen, Dipesh Tamboli, Tian Lan, Vaneet Aggarwal

Multi-task Imitation Learning (MIL) aims to train a policy capable of performing a distribution of tasks based on multi-task expert demonstrations, which is essential for general-purpose robots. Existing MIL algorithms suffer from low data efficiency and poor performance on complex long-horizontal tasks. We develop Multi-task Hierarchical Adversarial Inverse Reinforcement Learning (MH-AIRL) to learn hierarchically-structured multi-task policies, which is more beneficial for compositional tasks with long horizons and has higher expert data efficiency through identifying and transferring reusable basic skills across tasks. To realize this, MH-AIRL effectively synthesizes context-based multi-task learning, AIRL (an IL approach), and hierarchical policy learning. Further, MH-AIRL can be adopted to demonstrations without the task or skill annotations (i.e., state-action pairs only) which are more accessible in practice. Theoretical justifications are provided for each module of MH-AIRL, and evaluations on challenging multi-task settings demonstrate superior performance and transferability of the multi-task policies learned with MH-AIRL as compared to SOTA MIL baselines.

****

Model Transferability with Responsive Decision Subjects

Yatong Chen, Zeyu Tang, Kun Zhang, Yang Liu

Given an algorithmic predictor that is accurate on some source population consisting of strategic human decision subjects, will it remain accurate if the population respond to it? In our setting, an agent or a user corresponds to a sample (X,Y) drawn from a distribution $\cal{D}$ and will face a model h and its classification result $h(X)$. Agents can modify X to adapt to h, which will incur a distribution shift on (X,Y). Our formulation is motivated by applications where the deployed machine learning models are subjected to human agents, and will ultimately face responsive and interactive data distributions. We formalize the discussions of the transferability of a model by studying how the performance of the model trained on the available source distribution (data) would translate to the performance on its induced domain. We provide both upper bounds for the performance gap due to the induced domain shift, as well as lower bounds for the trade-offs that a classifier has to suffer on either the source training distribution or the induced target distribution. We provide further instantiated analysis for two popular domain adaptation settings, including covariate shift and target shift.

****

Layered State Discovery for Incremental Autonomous Exploration

Liyu Chen, Andrea Tirinzoni, Alessandro Lazaric, Matteo Pirotta

We study the autonomous exploration (AX) problem proposed by Lim & Auer (2012). In this setting, the objective is to discover a set of ϵ-optimal policies reaching a set $\mathcal{S}_L^{\rightarrow}$ of incrementally L-controllable states. We introduce a novel layered decomposition of the set of incrementally L-controllable states that is based on the iterative application of a state-expansion operator. We leverage these results to design Layered Autonomous Exploration (LAE), a novel algorithm for AX that attains a sample complexity of $\tilde{\mathcal{O}}(LS^{\rightarrow}_{L(1+\epsilon)}\Gamma_{L(1+\epsilon)} A \ln^{12}(S^{\rightarrow}_{L(1+\epsilon)})/\epsilon^2)$, where $S^{\rightarrow}_{L(1+\epsilon)}$ is the number of states that are incrementally $L(1+\epsilon)$-controllable, A is the number of actions, and $\Gamma_{L(1+\epsilon)}$ is the branching factor of the transitions over such states. LAE improves over the algorithm of Tarbouriech et al. (2020a) by a factor of L^2 and it is the first algorithm for AX that works in a countably-infinite state space. Moreover, we show that, under a certain identifiability assumption, LAE achieves minimax-optimal sample com

plexity of $\tilde{\mathcal{O}}(LS^{\rightarrow}_{L}A\ln^{12}(S^{\rightarrow}_{L})/\epsilon^2)$, outperforming existing algorithms and matching for the first time the lower bound proved by Cai et al. (2022) up to logarithmic factors.

****

Optimistic Online Mirror Descent for Bridging Stochastic and Adversarial Online Convex Optimization

Sijia Chen, Wei-Wei Tu, Peng Zhao, Lijun Zhang

Stochastically Extended Adversarial (SEA) model is introduced by Sachs et al. (2022) as an interpolation between stochastic and adversarial online convex optimization. Under the smoothness condition, they demonstrate that the expected regret of optimistic follow-the-regularized-leader (FTRL) depends on the cumulative stochastic variance $\sigma_{1:T}^2$ and the cumulative adversarial variation $\Sigma_{1:T}^2$ for convex functions. They also provide a slightly weaker bound based on the maximal stochastic variance $\sigma_{\max}^2$ and the maximal adversarial variation $\Sigma_{\max}^2$ for strongly convex functions. Inspired by their work, we investigate the theoretical guarantees of optimistic online mirror descent (OMD) for the SEA model. For convex and smooth functions, we obtain the same $\mathcal{O}(\sqrt{\sigma_{1:T}^2}+\sqrt{\Sigma_{1:T}^2})$ regret bound, without the convexity requirement of individual functions. For strongly convex and smooth functions, we establish an $\mathcal{O}(\min\{\log (\sigma_{1:T}^2+\Sigma_{1:T}^2), (\sigma_{\max}^2 + \Sigma_{\max}^2) \log T\})$ bound, better than their $\mathcal{O}((\sigma_{\max}^2 + \Sigma_{\max}^2) \log T)$ result. For exp-concave and smooth functions, we achieve a new $\mathcal{O}(d\log(\sigma_{1:T}^2+\Sigma_{1:T}^2))$ bound. Owing to the OMD framework, we further establish dynamic regret for convex and smooth functions, which is more favorable in non-stationary online scenarios.

****

Learning to Optimize Differentiable Games

Xuxi Chen, Nelson Vadori, Tianlong Chen, Zhangyang Wang

Many machine learning problems can be abstracted in solving game theory formulations and boil down to optimizing nested objectives, such as generative adversarial networks (GANs) and multi-agent reinforcement learning. Solving these games requires finding their stable fixed points or Nash equilibrium. However, existing algorithms for solving games suffer from empirical instability, hence demanding heavy ad-hoc tuning in practice. To tackle these challenges, we resort to the emerging scheme of Learning to Optimize (L2O), which discovers problem-specific efficient optimization algorithms through data-driven training. Our customized L2O framework for differentiable game theory problems, dubbed “Learning to Play Games" (L2PG), seeks a stable fixed point solution, by predicting the fast update direction from the past trajectory, with a novel gradient stability-aware, sign-based loss function. We further incorporate curriculum learning and self-learning to strengthen the empirical training stability and generalization of L2PG. On test problems including quadratic games and GANs, L2PG can substantially accelerate the convergence, and demonstrates a remarkably more stable trajectory. Codes are available at https://github.com/VITA-Group/L2PG.

****

Coordinated Dynamic Bidding in Repeated Second-Price Auctions with Budgets

Yurong Chen, Qian Wang, Zhijian Duan, Haoran Sun, Zhaohua Chen, Xiang Yan, Xiaotie Deng

In online ad markets, a rising number of advertisers are employing bidding agencies to participate in ad auctions. These agencies are specialized in designing online algorithms and bidding on behalf of their clients. Typically, an agency usually has information on multiple advertisers, so she can potentially coordinate bids to help her clients achieve higher utilities than those under independent bidding. In this paper, we study coordinated online bidding algorithms in repeated second-price auctions with budgets. We propose algorithms that guarantee every client a higher utility than the best she can get under independent bidding. We show that these algorithms achieve maximal social welfare and discuss bidders’ incentives to misreport their budgets, in symmetric cases. Our proofs combine the techniques of online learning and equilibrium analysis, overcoming the diffic

ulty of competing with a multi-dimensional benchmark. The performance of our alg orithms is further evaluated by experiments on both synthetic and real data. To the best of our knowledge, we are the first to consider bidder coordination in o nline repeated auctions with constraints.
****
Semi-Offline Reinforcement Learning for Optimized Text Generation
Changyu Chen, Xiting Wang, Yiqiao Jin, Victor Ye Dong, Li Dong, Jie Cao, Yi Liu, Rui Yan
Existing reinforcement learning (RL) mainly utilize online or offline settings. The online methods explore the environment with expensive time cost, and the off line methods efficiently obtain reward signals by sacrificing the exploration ca pability. We propose semi-offline RL, a novel paradigm that can smoothly transit from the offline setting to the online setting, balances the exploration capabi lity and training cost, and provides a theoretical foundation for comparing diff erent RL settings. Based on the semi-offline MDP formulation, we present the RL setting that is optimal in terms of optimization cost, asymptotic error, and ove rfitting error bound. Extensive experiments show that our semi-offline RL approa ch is effective in various text generation tasks and datasets, and yields compar able or usually better performance compared with the state-of-the-art methods.
****
Lower Bounds for Learning in Revealing POMDPs
Fan Chen, Huan Wang, Caiming Xiong, Song Mei, Yu Bai
This paper studies the fundamental limits of reinforcement learning (RL) in the challenging partially observable setting. While it is well-established that lear ning in Partially Observable Markov Decision Processes (POMDPs) requires exponen tially many samples in the worst case, a surge of recent work shows that polynom ial sample complexities are achievable under the revealing condition—A natural c ondition that requires the observables to reveal some information about the unob served latent states. However, the fundamental limits for learning in revealing POMDPs are much less understood, with existing lower bounds being rather prelimi nary and having substantial gaps from the current best upper bounds. We establis h strong PAC and regret lower bounds for learning in revealing POMDPs. Our lower bounds scale polynomially in all relevant problem parameters in a multiplicativ e fashion, and achieve significantly smaller gaps against the current best upper bounds, providing a solid starting point for future studies. In particular, for multi-step revealing POMDPs, we show that (1) the latent state-space dependence is at least $\Omega(S^{1.5})$ in the PAC sample complexity, which is notably ha rder than the $\widetilde{\Theta}(S)$ scaling for fully-observable MDPs; (2) Any polynomial sublinear regret is at least $\Omega(T^{2/3})$, suggesting its funda mental difference from the single-step case where $\widetilde{\mathcal{O}}(\sqrt {T})$ regret is achievable. Technically, our hard instance construction adapts t echniques in distribution testing, which is new to the RL literature and may be of independent interest. We also complement our results with new sharp regret up per bounds for strongly B-stable PSRs, which include single-step revealing POMDP s as a special case.
****
Implicit Neural Spatial Representations for Time-dependent PDEs
Honglin Chen, Rundi Wu, Eitan Grinspun, Changxi Zheng, Peter Yichen Chen
Implicit Neural Spatial Representation (INSR) has emerged as an effective repres entation of spatially-dependent vector fields. This work explores solving time-d ependent PDEs with INSR. Classical PDE solvers introduce both temporal and spati al discretizations. Common spatial discretizations include meshes and meshless p oint clouds, where each degree-of-freedom corresponds to a location in space. Wh ile these explicit spatial correspondences are intuitive to model and understand , these representations are not necessarily optimal for accuracy, memory usage, or adaptivity. Keeping the classical temporal discretization unchanged (e.g., ex plicit/implicit Euler), we explore INSR as an alternative spatial discretization , where spatial information is implicitly stored in the neural network weights. The network weights then evolve over time via time integration. Our approach doe s not require any training data generated by existing solvers because our approa

ch is the solver itself. We validate our approach on various PDEs with examples involving large elastic deformations, turbulent fluids, and multi-scale phenomena. While slower to compute than traditional representations, our approach exhibits higher accuracy and lower memory consumption. Whereas classical solvers can dynamically adapt their spatial representation only by resorting to complex remeshing algorithms, our INSR approach is intrinsically adaptive. By tapping into the rich literature of classic time integrators, e.g., operator-splitting schemes, our method enables challenging simulations in contact mechanics and turbulent flows where previous neural-physics approaches struggle. Videos and codes are available on the project page: http://www.cs.columbia.edu/cg/INSR-PDE/
****

BEATs: Audio Pre-Training with Acoustic Tokenizers
Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, Furu Wei
We introduce a self-supervised learning (SSL) framework BEATs for general audio representation pre-training, where we optimize an acoustic tokenizer and an audio SSL model by iterations. Unlike the previous audio SSL models that employ reconstruction loss for pre-training, our audio SSL model is trained with the discrete label prediction task, where the labels are generated by a semantic-rich acoustic tokenizer. We propose an iterative pipeline to jointly optimize the tokenizer and the pre-trained model, aiming to abstract high-level semantics and discard the redundant details for audio. The experimental results demonstrate our acoustic tokenizers can generate discrete labels with rich audio semantics and our audio SSL models achieve state-of-the-art (SOTA) results across various audio classification benchmarks, even outperforming previous models that use more training data and model parameters significantly. Specifically, we set a new SOTA mAP 50.6% on AudioSet-2M without using any external data, and 98.1% accuracy on ESC-50. The code and pre-trained models are available at https://aka.ms/beats.
****

Learning to Incentivize Information Acquisition: Proper Scoring Rules Meet Principal-Agent Model
Siyu Chen, Jibang Wu, Yifan Wu, Zhuoran Yang
We study the incentivized information acquisition problem, where a principal hires an agent to gather information on her behalf. Such a problem is modeled as a Stackelberg game between the principal and the agent, where the principal announces a scoring rule that specifies the payment, and then the agent then chooses an effort level that maximizes her own profit and reports the information. We study the online setting of such a problem from the principal’s perspective, i.e., designing the optimal scoring rule by repeatedly interacting with the strategic agent. We design a provably sample efficient algorithm that tailors the UCB algorithm (Auer et al., 2002) to our model, which achieves a $\mathcal{O} (K^2\cdot T^{2/3})$ regret after T iterations, where K is the number of effort levels of the agent. Our algorithm features a delicate estimation procedure for the optimal profit of the principal, and a conservative correction scheme that ensures the desired agent’s actions are incentivized. Furthermore, a key feature of our regret bound is that it is independent of the number of states of the environment.
****

Faster Gradient-Free Algorithms for Nonsmooth Nonconvex Stochastic Optimization
Lesi Chen, Jing Xu, Luo Luo
We consider the optimization problem of the form $\min_{x \in \mathbb{R}^d} f(x) \triangleq \mathbb{E}[F(x;\xi)]$, where the component $F(x;\xi)$ is L-mean-squared Lipschitz but possibly nonconvex and nonsmooth.The recently proposed gradient-free method requires at most $\mathcal{O}(L^4 d^{3/2} \epsilon^{-4} + \Delta L^3 d^{3/2} \delta^{-1} \epsilon^{-4})$ stochastic zeroth-order oracle complexity to find a (δ,ϵ)-Goldstein stationary point of objective function, where $\Delta = f(x_0) - \inf_{x \in \mathbb{R}^d} f(x)$ and x_0 is the initial point of the algorithm. This paper proposes a more efficient algorithm using stochastic recursive gradient estimators, which improves the complexity to $\mathcal{O}(L^3 d^{3/2} \epsilon^{-3}+ \Delta L^2 d^{3/2} \delta^{-1} \epsilon^{

-3})$.
****
Efficient Personalized Federated Learning via Sparse Model-Adaptation
Daoyuan Chen, Liuyi Yao, Dawei Gao, Bolin Ding, Yaliang Li
Federated Learning (FL) aims to train machine learning models for multiple clien ts without sharing their own private data. Due to the heterogeneity of clients' local data distribution, recent studies explore the personalized FL that learns and deploys distinct local models with the help of auxiliary global models. Howe ver, the clients can be heterogeneous in terms of not only local data distributi on, but also their computation and communication resources. The capacity and eff iciency of personalized models are restricted by the lowest-resource clients, le ading to sub-optimal performance and limited practicality of personalized FL. To overcome these challenges, we propose a novel approach named pFedGate for effic ient personalized FL by adaptively and efficiently learning sparse local models. With a lightweight trainable gating layer, pFedGate enables clients to reach th eir full potential in model capacity by generating different sparse models accou nting for both the heterogeneous data distributions and resource constraints. Me anwhile, the computation and communication efficiency are both improved thanks t o the adaptability between the model sparsity and clients' resources. Further, w e theoretically show that the proposed pFedGate has superior complexity with gua ranteed convergence and generalization error. Extensive experiments show that pF edGate achieves superior global accuracy, individual accuracy and efficiency sim ultaneously over state-of-the-art methods. We also demonstrate that pFedGate per forms better than competitors in the novel clients participation and partial cli ents participation scenarios, and can learn meaningful sparse local models adapt ed to different data distributions.
****
A Gromov-Wasserstein Geometric View of Spectrum-Preserving Graph Coarsening
Yifan Chen, Rentian Yao, Yun Yang, Jie Chen
Graph coarsening is a technique for solving large-scale graph problems by workin g on a smaller version of the original graph, and possibly interpolating the res ults back to the original graph. It has a long history in scientific computing a nd has recently gained popularity in machine learning, particularly in methods t hat preserve the graph spectrum. This work studies graph coarsening from a diffe rent perspective, developing a theory for preserving graph distances and proposi ng a method to achieve this. The geometric approach is useful when working with a collection of graphs, such as in graph classification and regression. In this study, we consider a graph as an element on a metric space equipped with the Gro mov–Wasserstein (GW) distance, and bound the difference between the distance of two graphs and their coarsened versions. Minimizing this difference can be done using the popular weighted kernel K-means method, which improves existing spec trum-preserving methods with the proper choice of the kernel. The study includes a set of experiments to support the theory and method, including approximating the GW distance, preserving the graph spectrum, classifying graphs using spectra l information, and performing regression using graph convolutional networks. Cod e is available at https://github.com/ychen-stat-ml/GW-Graph-Coarsening.
****
How to address monotonicity for model risk management?
Dangxing Chen, Weicheng Ye
In this paper, we study the problem of establishing the accountability and fairn ess of transparent machine learning models through monotonicity. Although there have been numerous studies on individual monotonicity, pairwise monotonicity is often overlooked in the existing literature. This paper studies transparent neur al networks in the presence of three types of monotonicity: individual monotonic ity, weak pairwise monotonicity, and strong pairwise monotonicity. As a means of achieving monotonicity while maintaining transparency, we propose the monotonic groves of neural additive models. As a result of empirical examples, we demonst rate that monotonicity is often violated in practice and that monotonic groves o f neural additive models are transparent, accountable, and fair.
****

Sketched Ridgeless Linear Regression: The Role of Downsampling
Xin Chen, Yicheng Zeng, Siyue Yang, Qiang Sun
Overparametrization often helps improve the generalization performance. This pap er presents a dual view of overparametrization suggesting that downsampling may also help generalize. Focusing on the proportional regime $m\asymp n \asymp p$, where m represents the sketching size, n is the sample size, and p is the feature dimensionality, we investigate two out-of-sample prediction risks of the sketched ridgeless least square estimator. Our findings challenge conventional beliefs by showing that downsampling does not always harm generalization but can actually improve it in certain cases. We identify the optimal sketching size th at minimizes out-of-sample prediction risks and demonstrate that the optimally s ketched estimator exhibits stabler risk curves, eliminating the peaks of those f or the full-sample estimator. To facilitate practical implementation, we propose an empirical procedure to determine the optimal sketching size. Finally, we ext end our analysis to cover central limit theorems and misspecified models. Numeri cal studies strongly support our theory.
****
Context-Aware Bayesian Network Actor-Critic Methods for Cooperative Multi-Agent Reinforcement Learning
Dingyang Chen, Qi Zhang
Executing actions in a correlated manner is a common strategy for human coordina tion that often leads to better cooperation, which is also potentially beneficia l for cooperative multi-agent reinforcement learning (MARL). However, the recent success of MARL relies heavily on the convenient paradigm of purely decentraliz ed execution, where there is no action correlation among agents for scalability considerations. In this work, we introduce a Bayesian network to inaugurate corr elations between agents' action selections in their joint policy. Theoretically, we establish a theoretical justification for why action dependencies are benefi cial by deriving the multi-agent policy gradient formula under such a Bayesian n etwork joint policy and proving its global convergence to Nash equilibria under tabular softmax policy parameterization in cooperative Markov games. Further, by equipping existing MARL algorithms with a recent method of differentiable direc ted acyclic graphs (DAGs), we develop practical algorithms to learn the context- aware Bayesian network policies in scenarios with partial observability and vari ous difficulty. We also dynamically decrease the sparsity of the learned DAG thr oughout the training process, which leads to weakly or even purely independent p olicies for decentralized execution. Empirical results on a range of MARL benchm arks show the benefits of our approach.
****
Bidirectional Learning for Offline Model-based Biological Sequence Design
Can Chen, Yingxue Zhang, Xue Liu, Mark Coates
Offline model-based optimization aims to maximize a black-box objective function with a static dataset of designs and their scores. In this paper, we focus on b iological sequence design to maximize some sequence score. A recent approach emp loys bidirectional learning, combining a forward mapping for exploitation and a backward mapping for constraint, and it relies on the neural tangent kernel (NTK) of an infinitely wide network to build a proxy model. Though effective, the NT K cannot learn features because of its parametrization, and its use prevents the incorporation of powerful pre-trained Language Models (LMs) that can capture th e rich biophysical information in millions of biological sequences. We adopt an alternative proxy model, adding a linear head to a pre-trained LM, and propose a linearization scheme. This yields a closed-form loss and also takes into accoun t the biophysical information in the pre-trained LM. In addition, the forward ma pping and the backward mapping play different roles and thus deserve different w eights during sequence optimization. To achieve this, we train an auxiliary mode l and leverage its weak supervision signal via a bi-level optimization framework to effectively learn how to balance the two mappings. Further, by extending the framework, we develop the first learning rate adaptation module Adaptive-η , which is compatible with all gradient-based algorithms for offline model-based optimization. Experimental results on DNA/protein sequence design tasks verify

the effectiveness of our algorithm. Our code is available at https://github.com/GGchen1997/BIB-ICML2023-Submission.
****

Learning to Jump: Thinning and Thickening Latent Counts for Generative Modeling
Tianqi Chen, Mingyuan Zhou
Learning to denoise has emerged as a prominent paradigm to design state-of-the-art deep generative models for natural images. How to use it to model the distributions of both continuous real-valued data and categorical data has been well studied in recently proposed diffusion models. However, it is found in this paper to have limited ability in modeling some other types of data, such as count and non-negative continuous data, that are often highly sparse, skewed, heavy-tailed, and/or overdispersed. To this end, we propose learning to jump as a general recipe for generative modeling of various types of data. Using a forward count thinning process to construct learning objectives to train a deep neural network, it employs a reverse count thickening process to iteratively refine its generation through that network. We demonstrate when learning to jump is expected to perform comparably to learning to denoise, and when it is expected to perform better. For example, learning to jump is recommended when the training data is non-negative and exhibits strong sparsity, skewness, heavy-tailedness, and/or heterogeneity.
****

Lifelong Language Pretraining with Distribution-Specialized Experts
Wuyang Chen, Yanqi Zhou, Nan Du, Yanping Huang, James Laudon, Zhifeng Chen, Claire Cui
Pretraining on a large-scale corpus has become a standard method to build general language models (LMs). Adapting a model to new data distributions targeting different downstream tasks poses significant challenges. Naive fine-tuning may incur catastrophic forgetting when the over-parameterized LMs overfit the new data but fail to preserve the pretrained features. Lifelong learning (LLL) aims to enable information systems to learn from a continuous data stream across time. However, most prior work modifies the training recipe assuming a static fixed network architecture. We find that additional model capacity and proper regularization are key elements to achieving strong LLL performance. Thus, we propose Lifelong-MoE, an extensible MoE (Mixture-of-Experts) architecture that dynamically adds model capacity via adding experts with regularized pretaining. Our results show that by only introducing a limited number of extra experts while keeping the computation cost constant, our model can steadily adapt to data distribution shifts while preserving the previous knowledge. Compared to existing lifelong learning approaches, Lifelong-MoE achieves better few-shot performance on NLP tasks. More impressively, Lifelong-MoE surpasses multi-task learning on 19 downstream NLU tasks.
****

Generalized-Smooth Nonconvex Optimization is As Efficient As Smooth Nonconvex Optimization
Ziyi Chen, Yi Zhou, Yingbin Liang, Zhaosong Lu
Various optimal gradient-based algorithms have been developed for smooth nonconvex optimization. However, many nonconvex machine learning problems do not belong to the class of smooth functions and therefore the existing algorithms are sub-optimal. Instead, these problems have been shown to satisfy certain generalized-smooth conditions, which have not been well understood in the existing literature. In this paper, we propose a notion of α-symmetric generalized-smoothness that substantially extends the existing notions and covers many important functions such as high-order polynomials and exponential functions. We study the fundamental properties and establish descent lemmas for the functions in this class. Then, to solve such a large class of nonconvex problems, we design a special deterministic normalized gradient descent algorithm that achieves the optimal iteration complexity $\mathcal{O}(\epsilon^{-2})$, and also prove that the popular SPIDER variance reduction algorithm achieves the optimal sample complexity $\mathcal{O}(\epsilon^{-3})$. Our results show that solving generalized-smooth nonconvex problems is as efficient as solving smooth nonconvex problems.

****
Weakly Supervised Regression with Interval Targets
Xin Cheng, Yuzhou Cao, Ximing Li, Bo An, Lei Feng
This paper investigates an interesting weakly supervised regression setting called regression with interval targets (RIT). Although some of the previous methods on relevant regression settings can be adapted to RIT, they are not statistically consistent, and thus their empirical performance is not guaranteed. In this paper, we provide a thorough study on RIT. First, we proposed a novel statistical model to describe the data generation process for RIT and demonstrate its validity. Second, we analyze a simple selecting method for RIT, which selects a particular value in the interval as the target value to train the model. Third, we propose a statistically consistent limiting method for RIT to train the model by limiting the predictions to the interval. We further derive an estimation error bound for our limiting method. Finally, extensive experiments on various datasets demonstrate the effectiveness of our proposed method.
****
PLay: Parametrically Conditioned Layout Generation using Latent Diffusion
Chin-Yi Cheng, Forrest Huang, Gang Li, Yang Li
Layout design is an important task in various design fields, including user interfaces, document, and graphic design. As this task requires tedious manual effort by designers, prior works have attempted to automate this process using generative models, but commonly fell short of providing intuitive user controls and achieving design objectives. In this paper, we build a conditional latent diffusion model, PLay, that generates parametrically conditioned layouts in vector graphic space from user-specified guidelines, which are commonly used by designers for representing their design intents in current practices. Our method outperforms prior works across three datasets on metrics including FID and FD-VG, and in user test. Moreover, it brings a novel and interactive experience to professional layout design processes.
****
Identification of the Adversary from a Single Adversarial Example
Minhao Cheng, Rui Min, Haochen Sun, Pin-Yu Chen
Deep neural networks have been shown vulnerable to adversarial examples. Even though many defense methods have been proposed to enhance the robustness, it is still a long way toward providing an attack-free method to build a trustworthy machine learning system. In this paper, instead of enhancing the robustness, we take the investigator’s perspective and propose a new framework to trace the first compromised model copy in a forensic investigation manner. Specifically, we focus on the following setting: the machine learning service provider provides model copies for a set of customers. However, one of the customers conducted adversarial attacks to fool the system. Therefore, the investigator’s objective is to identify the first compromised copy by collecting and analyzing evidence from only available adversarial examples. To make the tracing viable, we design a random mask watermarking mechanism to differentiate adversarial examples from different copies. First, we propose a tracing approach in the data-limited case where the original example is also available. Then, we design a data-free approach to identify the adversary without accessing the original example. Finally, the effectiveness of our proposed framework is evaluated by extensive experiments with different model architectures, adversarial attacks, and datasets.
****
Parallel Online Clustering of Bandits via Hedonic Game
Xiaotong Cheng, Cheng Pan, Setareh Maghsudi
Contextual bandit algorithms appear in several applications, such as online advertisement and recommendation systems like personalized education or personalized medicine. Individually-tailored recommendations boost the performance of the underlying application; nevertheless, providing individual suggestions becomes costly and even implausible as the number of users grows. As such, to efficiently serve the demands of several users in modern applications, it is imperative to identify the underlying users’ clusters, i.e., the groups of users for which a single recommendation might be (near-)optimal. We propose CLUB-HG, a novel algorith

m that integrates a game-theoretic approach into clustering inference. Our algor
ithm achieves Nash equilibrium at each inference step and discovers the underlyi
ng clusters. We also provide regret analysis within a standard linear stochastic
 noise setting. Finally, experiments on synthetic and real-world datasets show t
he superior performance of our proposed algorithm compared to the state-of-the-a
rt algorithms.
****

Mu^{2}SLAM: Multitask, Multilingual Speech and Language Models
Yong Cheng, Yu Zhang, Melvin Johnson, Wolfgang Macherey, Ankur Bapna
We present Mu^{2}SLAM, a multilingual sequence-to-sequence model pre-trained joi
ntly on unlabeled speech, unlabeled text and supervised data spanning Automatic
Speech Recognition (ASR), Automatic Speech Translation (AST) and Machine Transla
tion (MT), in over 100 languages. By leveraging a quantized representation of sp
eech as a target, Mu^{2}SLAM trains the speech-text models with a sequence-to-se
quence masked denoising objective similar to T5 on the decoder and a masked lang
uage modeling objective (MLM) on the encoder, for both unlabeled speech and text
, while utilizing the supervised tasks to improve cross-lingual and cross-modal
representation alignment within the model. On CoVoST AST, Mu^{2}SLAM establishes
 a new state-of-the-art for models trained on public datasets, improving on xx-e
n translation over the previous best by 1.9 BLEU points and on en-xx translation
 by 1.1 BLEU points. On Voxpopuli ASR, our model matches the performance of an m
SLAM model fine-tuned with an RNN-T decoder, despite using a relatively weaker T
ransformer decoder. On text understanding tasks, our model improves by more than
 6% over mSLAM on XNLI, getting closer to the performance of mT5 models of compa
rable capacity on XNLI and TydiQA, paving the way towards a single model for all
 speech and text understanding tasks.
****

Understanding the Role of Feedback in Online Learning with Switching Costs
Duo Cheng, Xingyu Zhou, Bo Ji
In this paper, we study the role of feedback in online learning with switching c
osts. It has been shown that the minimax regret is $\widetilde{\Theta}(T^{2/3})$
 under bandit feedback and improves to $\widetilde{\Theta}(\sqrt{T})$ under full
-information feedback, where T is the length of the time horizon. However, it
remains largely unknown how the amount and type of feedback generally impact reg
ret. To this end, we first consider the setting of bandit learning with extra ob
servations; that is, in addition to the typical bandit feedback, the learner can
 freely make a total of B_{ex} extra observations. We fully character
ize the minimax regret in this setting, which exhibits an interesting phase-tran
sition phenomenon: when $B_{\mathrm{ex}} = O(T^{2/3})$, the regret remains $\wid
etilde{\Theta}(T^{2/3})$, but when $B_{\mathrm{ex}} = \Omega(T^{2/3})$, it becom
es $\widetilde{\Theta}(T/\sqrt{B_{\mathrm{ex}}})$, which improves as the budget
B_{ex} increases. To design algorithms that can achieve the minimax r
egret, it is instructive to consider a more general setting where the learner ha
s a budget of B total observations. We fully characterize the minimax regret i
n this setting as well and show that it is $\widetilde{\Theta}(T/\sqrt{B})$, whi
ch scales smoothly with the total budget B. Furthermore, we propose a generic
algorithmic framework, which enables us to design different learning algorithms
that can achieve matching upper bounds for both settings based on the amount and
 type of feedback. One interesting finding is that while bandit feedback can sti
ll guarantee optimal regret when the budget is relatively limited, it no longer
suffices to achieve optimal regret when the budget is relatively large.
****

Tighter Bounds on the Expressivity of Transformer Encoders
David Chiang, Peter Cholak, Anand Pillay
Characterizing neural networks in terms of better-understood formal systems has
the potential to yield new insights into the power and limitations of these netw
orks. Doing so for transformers remains an active area of research. Bhattamishra
 and others have shown that transformer encoders are at least as expressive as a
 certain kind of counter machine, while Merrill and Sabharwal have shown that fi
xed-precision transformer encoders recognize only languages in uniform TC^0. W

e connect and strengthen these results by identifying a variant of first-order l ogic with counting quantifiers that is simultaneously an upper bound for fixed-p recision transformer encoders and a lower bound for transformer encoders. This b rings us much closer than before to an exact characterization of the languages t hat transformer encoders recognize.
****

Provably Learning Diverse Features in Multi-View Data with Midpoint Mixup
Muthu Chidambaram, Xiang Wang, Chenwei Wu, Rong Ge
Mixup is a data augmentation technique that relies on training using random conv ex combinations of data points and their labels. In recent years, Mixup has beco me a standard primitive used in the training of state-of-the-art image classific ation models due to its demonstrated benefits over empirical risk minimization w ith regards to generalization and robustness. In this work, we try to explain so me of this success from a feature learning perspective. We focus our attention o n classification problems in which each class may have multiple associated featu res (or $\textit{views}$) that can be used to predict the class correctly. Our m ain theoretical results demonstrate that, for a non-trivial class of data distri butions with two features per class, training a 2-layer convolutional network us ing empirical risk minimization can lead to learning only one feature for almost all classes while training with a specific instantiation of Mixup succeeds in l earning both features for every class. We also show empirically that these theor etical insights extend to the practical settings of image benchmarks modified to have multiple features.
****

Hiding Data Helps: On the Benefits of Masking for Sparse Coding
Muthu Chidambaram, Chenwei Wu, Yu Cheng, Rong Ge
Sparse coding, which refers to modeling a signal as sparse linear combinations o f the elements of a learned dictionary, has proven to be a successful (and inter pretable) approach in applications such as signal processing, computer vision, a nd medical imaging. While this success has spurred much work on provable guarant ees for dictionary recovery when the learned dictionary is the same size as the ground-truth dictionary, work on the setting where the learned dictionary is lar ger (or $\textit{over-realized}$) with respect to the ground truth is comparativ ely nascent. Existing theoretical results in this setting have been constrained to the case of noise-less data. We show in this work that, in the presence of no ise, minimizing the standard dictionary learning objective can fail to recover t he elements of the ground-truth dictionary in the over-realized regime, regardle ss of the magnitude of the signal in the data-generating process. Furthermore, d rawing from the growing body of work on self-supervised learning, we propose a n ovel masking objective for which recovering the ground-truth dictionary is in fa ct optimal as the signal increases for a large class of data-generating processe s. We corroborate our theoretical results with experiments across several parame ter regimes showing that our proposed objective also enjoys better empirical per formance than the standard reconstruction objective.
****

PINA: Leveraging Side Information in eXtreme Multi-label Classification via Pred icted Instance Neighborhood Aggregation
Eli Chien, Jiong Zhang, Cho-Jui Hsieh, Jyun-Yu Jiang, Wei-Cheng Chang, Olgica Mi lenkovic, Hsiang-Fu Yu
The eXtreme Multi-label Classification (XMC) problem seeks to find relevant labe ls from an exceptionally large label space. Most of the existing XMC learners fo cus on the extraction of semantic features from input query text. However, conve ntional XMC studies usually neglect the side information of instances and labels , which can be of use in many real-world applications such as recommendation sys tems and e-commerce product search. We propose Predicted Instance Neighborhood A ggregation (PINA), a data augmentation method for the general XMC problem that l everages beneficial side information. Unlike most existing XMC frameworks that t reat labels and input instances as featureless indicators and independent entrie s, PINA extracts information from the label metadata and the correlations among training instances. Extensive experimental results demonstrate the consistent ga

in of PINA on various XMC tasks compared to the state-of-the-art methods: PINA o
ffers a gain in accuracy compared to standard XR-Transformers on five public ben
chmark datasets. Moreover, PINA achieves a ~ 5% gain in accuracy on the lar
gest dataset LF-AmazonTitles-1.3M.
****
Tight Certification of Adversarially Trained Neural Networks via Nonconvex Low-R
ank Semidefinite Relaxations
Hong-Ming Chiu, Richard Y. Zhang
Adversarial training is well-known to produce high-quality neural network models
 that are empirically robust against adversarial perturbations. Nevertheless, on
ce a model has been adversarially trained, one often desires a certification tha
t the model is truly robust against all future attacks. Unfortunately, when face
d with adversarially trained models, all existing approaches have significant tr
ouble making certifications that are strong enough to be practically useful. Lin
ear programming (LP) techniques in particular face a “convex relaxation barrier”
 that prevent them from making high-quality certifications, even after refinemen
t with mixed-integer linear programming (MILP) and branch-and-bound (BnB) techni
ques. In this paper, we propose a nonconvex certification technique, based on a
low-rank restriction of a semidefinite programming (SDP) relaxation. The nonconv
ex relaxation makes strong certifications comparable to much more expensive SDP
methods, while optimizing over dramatically fewer variables comparable to much w
eaker LP methods. Despite nonconvexity, we show how off-the-shelf local optimiza
tion algorithms can be used to achieve and to certify global optimality in polyn
omial time. Our experiments find that the nonconvex relaxation almost completely
 closes the gap towards exact certification of adversarially trained models.
****
Neural Latent Aligner: Cross-trial Alignment for Learning Representations of Com
plex, Naturalistic Neural Data
Cheol Jun Cho, Edward Chang, Gopala Anumanchipalli
Understanding the neural implementation of complex human behaviors is one of the
 major goals in neuroscience. To this end, it is crucial to find a true represen
tation of the neural data, which is challenging due to the high complexity of be
haviors and the low signal-to-ratio (SNR) of the signals. Here, we propose a nov
el unsupervised learning framework, Neural Latent Aligner (NLA), to find well-co
nstrained, behaviorally relevant neural representations of complex behaviors. Th
e key idea is to align representations across repeated trials to learn cross-tri
al consistent information. Furthermore, we propose a novel, fully differentiable
 time warping model (TWM) to resolve the temporal misalignment of trials. When a
pplied to intracranial electrocorticography (ECoG) of natural speaking, our mode
l learns better representations for decoding behaviors than the baseline models,
 especially in lower dimensional space. The TWM is empirically validated by meas
uring behavioral coherence between aligned trials. The proposed framework learns
 more cross-trial consistent representations than the baselines, and when visual
ized, the manifold reveals shared neural trajectories across trials.
****
On the Convergence of Federated Averaging with Cyclic Client Participation
Yae Jee Cho, Pranay Sharma, Gauri Joshi, Zheng Xu, Satyen Kale, Tong Zhang
Federated Averaging (FedAvg) and its variants are the most popular optimization
algorithms in federated learning (FL). Previous convergence analyses of FedAvg e
ither assume full client participation or partial client participation where the
 clients can be uniformly sampled. However, in practical cross-device FL systems
, only a subset of clients that satisfy local criteria such as battery status, n
etwork connectivity, and maximum participation frequency requirements (to ensure
 privacy) are available for training at a given time. As a result, client availa
bility follows a natural cyclic pattern. We provide (to our knowledge) the first
 theoretical framework to analyze the convergence of FedAvg with cyclic client p
articipation with several different client optimizers such as GD, SGD, and shuff
led SGD. Our analysis discovers that cyclic client participation can achieve a f
aster asymptotic convergence rate than vanilla FedAvg with uniform client partic
ipation under suitable conditions, providing valuable insights into the design o

f client sampling protocols.
****
GREAD: Graph Neural Reaction-Diffusion Networks
Jeongwhan Choi, Seoyoung Hong, Noseong Park, Sung-Bae Cho
Graph neural networks (GNNs) are one of the most popular research topics for dee p learning. GNN methods typically have been designed on top of the graph signal processing theory. In particular, diffusion equations have been widely used for designing the core processing layer of GNNs, and therefore they are inevitably v ulnerable to the notorious oversmoothing problem. Recently, a couple of papers p aid attention to reaction equations in conjunctions with diffusion equations. Ho wever, they all consider limited forms of reaction equations. To this end, we pr esent a reaction-diffusion equation-based GNN method that considers all popular types of reaction equations in addition to one special reaction equation designe d by us. To our knowledge, our paper is one of the most comprehensive studies on reaction-diffusion equation-based GNNs. In our experiments with 9 datasets and 28 baselines, our method, called GREAD, outperforms them in a majority of cases. Further synthetic data experiments show that it mitigates the oversmoothing pro blem and works well for various homophily rates.
****
Is Overfitting Necessary for Implicit Video Representation?
Hee Min Choi, Hyoa Kang, Dokwan Oh
Compact representation of multimedia signals using implicit neural representatio ns (INRs) has advanced significantly over the past few years, and recent works a ddress their applications to video. Existing studies on video INR have focused o n network architecture design as all video information is contained within netwo rk parameters. Here, we propose a new paradigm in efficient INR for videos based on the idea of strong lottery ticket (SLT) hypothesis (Zhou et al., 2019), whic h demonstrates the possibility of finding an accurate subnetwork mask, called su permask, for a randomly initialized classification network without weight traini ng. Specifically, we train multiple supermasks with a hierarchical structure for a randomly initialized image-wise video representation model without weight upd ates. Different from a previous approach employing hierarchical supermasks (Okos hi et al., 2022), a trainable scale parameter for each mask is used instead of m ultiplying by the same fixed scale for all levels. This simple modification wide ns the parameter search space to sufficiently explore various sparsity patterns, leading the proposed algorithm to find stronger subnetworks. Moreover, extensiv e experiments on popular UVG benchmark show that random subnetworks obtained fro m our framework achieve higher reconstruction and visual quality than fully trai ned models with similar encoding sizes. Our study is the first to demonstrate th e existence of SLTs in video INR models and propose an efficient method for find ing them.
****
Semi-Parametric Contextual Pricing Algorithm using Cox Proportional Hazards Mode l
Young-Geun Choi, Gi-Soo Kim, Yunseo Choi, Wooseong Cho, Myunghee Cho Paik, Min-H wan Oh
Contextual dynamic pricing is a problem of setting prices based on current conte xtual information and previous sales history to maximize revenue. A popular appr oach is to postulate a distribution of customer valuation as a function of conte xtual information and the baseline valuation. A semi-parametric setting, where t he context effect is parametric and the baseline is nonparametric, is of growing interest due to its flexibility. A challenge is that customer valuation is almo st never observable in practice and is instead type-I interval censored by the o ffered price. To address this challenge, we propose a novel semi-parametric cont extual pricing algorithm for stochastic contexts, called the epoch-based Cox pro portional hazards Contextual Pricing (CoxCP) algorithm. To our best knowledge, o ur work is the first to employ the Cox model for customer valuation. The CoxCP a lgorithm has a high-probability regret upper bound of $\tilde{O}(T^{\frac{2}{3} }d)$, where T is the length of horizon and d is the dimension of context. I n addition, if the baseline is known, the regret bound can improve to $O(d \log

T)$ under certain assumptions. We demonstrate empirically the proposed algorithm performs better than existing semi-parametric contextual pricing algorithms when the model assumptions of all algorithms are correct.
****

Restoration based Generative Models
Jaemoo Choi, Yesom Park, Myungjoo Kang
Denoising diffusion models (DDMs) have recently attracted increasing attention by showing impressive synthesis quality. DDMs are built on a diffusion process that pushes data to the noise distribution and the models learn to denoise. In this paper, we establish the interpretation of DDMs in terms of image restoration (IR). Integrating IR literature allows us to use an alternative objective and diverse forward processes, not confining to the diffusion process. By imposing prior knowledge on the loss function grounded on MAP-based estimation, we eliminate the need for the expensive sampling of DDMs. Also, we propose a multi-scale training, which improves the performance compared to the diffusion process, by taking advantage of the flexibility of the forward process. Experimental results demonstrate that our model improves the quality and efficiency of both training and inference. Furthermore, we show the applicability of our model to inverse problems. We believe that our framework paves the way for designing a new type of flexible general generative model.
****

Concept-based Explanations for Out-of-Distribution Detectors
Jihye Choi, Jayaram Raghuram, Ryan Feng, Jiefeng Chen, Somesh Jha, Atul Prakash
Out-of-distribution (OOD) detection plays a crucial role in ensuring the safe deployment of deep neural network (DNN) classifiers. While a myriad of methods have focused on improving the performance of OOD detectors, a critical gap remains in interpreting their decisions. We help bridge this gap by providing explanations for OOD detectors based on learned high-level concepts. We first propose two new metrics for assessing the effectiveness of a particular set of concepts for explaining OOD detectors: 1) detection completeness, which quantifies the sufficiency of concepts for explaining an OOD-detector’s decisions, and 2) concept separability, which captures the distributional separation between in-distribution and OOD data in the concept space. Based on these metrics, we propose an unsupervised framework for learning a set of concepts that satisfy the desired properties of high detection completeness and concept separability, and demonstrate its effectiveness in providing concept-based explanations for diverse off-the-shelf OOD detectors. We also show how to identify prominent concepts contributing to the detection results, and provide further reasoning about their decisions.
****

Active causal structure learning with advice
Davin Choo, Themistoklis Gouleakis, Arnab Bhattacharyya
We introduce the problem of active causal structure learning with advice. In the typical well-studied setting, the learning algorithm is given the essential graph for the observational distribution and is asked to recover the underlying causal directed acyclic graph (DAG) G^* while minimizing the number of interventions made. In our setting, we are additionally given side information about G^* as advice, e.g. a DAG G purported to be G^*. We ask whether the learning algorithm can benefit from the advice when it is close to being correct, while still having worst-case guarantees even when the advice is arbitrarily bad. Our work is in the same space as the growing body of research on algorithms with predictions. When the advice is a DAG G, we design an adaptive search algorithm to recover G^* whose intervention cost is at most $\mathcal{O}(\max\{1, \log \psi\})$ times the cost for verifying G^*; here, ψ is a distance measure between G and G^* that is upper bounded by the number of variables n, and is exactly 0 when $G=G^*$. Our approximation factor matches the state-of-the-art for the advice-less setting.
****

New metrics and search algorithms for weighted causal DAGs
Davin Choo, Kirankumar Shiragur
Recovering causal relationships from data is an important problem. Using observa

tional data, one can typically only recover causal graphs up to a Markov equival ence class and additional assumptions or interventional data are needed for comp lete recovery. In this work, under some standard assumptions, we study causal gr aph discovery via adaptive interventions with node-dependent interventional cost s. For this setting, we show that no algorithm can achieve an approximation guar antee that is asymptotically better than linear in the number of vertices with r espect to the verification number; a well-established benchmark for adaptive sea rch algorithms. Motivated by this negative result, we define a new benchmark tha t captures the worst-case interventional cost for any search algorithm. Furtherm ore, with respect to this new benchmark, we provide adaptive search algorithms t hat achieve logarithmic approximations under various settings: atomic, bounded s ize interventions and generalized cost objectives.
****

Computational Doob h-transforms for Online Filtering of Discretely Observed Diff usions
Nicolas Chopin, Andras Fulop, Jeremy Heng, Alexandre H. Thiery
This paper is concerned with online filtering of discretely observed nonlinear d iffusion processes. Our approach is based on the fully adapted auxiliary particl e filter, which involves Doob's h-transforms that are typically intractable. W e propose a computational framework to approximate these h-transforms by solvi ng the underlying backward Kolmogorov equations using nonlinear Feynman-Kac form ulas and neural networks. The methodology allows one to train a locally optimal particle filter prior to the data-assimilation procedure. Numerical experiments illustrate that the proposed approach can be orders of magnitude more efficient than state-of-the-art particle filters in the regime of highly informative obser vations, when the observations are extreme under the model, and if the state dim ension is large.
****

Multi-Epoch Matrix Factorization Mechanisms for Private Machine Learning
Christopher A. Choquette-Choo, Hugh Brendan Mcmahan, J Keith Rush, Abhradeep Guh a Thakurta
We introduce new differentially private (DP) mechanisms for gradient-based machi ne learning (ML) with multiple passes (epochs) over a dataset, substantially imp roving the achievable privacy-utility-computation tradeoffs. We formalize the pr oblem of DP mechanisms for adaptive streams with multiple participations and int roduce a non-trivial extension of online matrix factorization DP mechanisms to o ur setting. This includes establishing the necessary theory for sensitivity calc ulations and efficient computation of optimal matrices. For some applications li ke $>\!\! 10,000$ SGD steps, applying these optimal techniques becomes computati onally expensive. We thus design an efficient Fourier-transform-based mechanism with only a minor utility loss. Extensive empirical evaluation on both example-l evel DP for image classification and user-level DP for language modeling demonst rate substantial improvements over all previous methods, including the widely-us ed DP-SGD. Though our primary application is to ML, our main DP results are appl icable to arbitrary linear queries and hence may have much broader applicability .
****

Taming graph kernels with random features
Krzysztof Marcin Choromanski
We introduce in this paper the mechanism of graph random features (GRFs). GRFs c an be used to construct unbiased randomized estimators of several important kern els defined on graphs' nodes, in particular the regularized Laplacian kernel. As regular RFs for non-graph kernels, they provide means to scale up kernel method s defined on graphs to larger networks. Importantly, they give substantial compu tational gains also for smaller graphs, while applied in downstream applications . Consequently, GRFs address the notoriously difficult problem of cubic (in the number of the nodes of the graph) time complexity of graph kernels algorithms. W e provide a detailed theoretical analysis of GRFs and an extensive empirical eva luation: from speed tests, through Frobenius relative error analysis to kmeans g raph-clustering with graph kernels. We show that the computation of GRFs admits

an embarrassingly simple distributed algorithm that can be applied if the graph under consideration needs to be split across several machines. We also introduce a (still unbiased) quasi Monte Carlo variant of GRFs, q-GRFs, relying on the so-called reinforced random walks that might be used to optimize the variance of GRFs. As a byproduct, we obtain a novel approach to solve certain classes of linear equations with positive and symmetric matrices.
****
Efficient Graph Field Integrators Meet Point Clouds
Krzysztof Marcin Choromanski, Arijit Sehanobish, Han Lin, Yunfan Zhao, Eli Berger, Tetiana Parshakova, Alvin Pan, David Watkins, Tianyi Zhang, Valerii Likhosherstov, Somnath Basu Roy Chowdhury, Kumar Avinava Dubey, Deepali Jain, Tamas Sarlos, Snigdha Chaturvedi, Adrian Weller
We present two new classes of algorithms for efficient field integration on graphs encoding point cloud data. The first class, $\mathrm{SeparatorFactorization}$ (SF), leverages the bounded genus of point cloud mesh graphs, while the second class, $\mathrm{RFDiffusion}$ (RFD), uses popular ϵ-nearest-neighbor graph representations for point clouds. Both can be viewed as providing the functionality of Fast Multipole Methods (FMMs), which have had a tremendous impact on efficient integration, but for non-Euclidean spaces. We focus on geometries induced by distributions of walk lengths between points (e.g. shortest-path distance). We provide an extensive theoretical analysis of our algorithms, obtaining new results in structural graph theory as a byproduct. We also perform exhaustive empirical evaluation, including on-surface interpolation for rigid and deformable objects (in particular for mesh-dynamics modeling) as well as Wasserstein distance computations for point clouds, including the Gromov-Wasserstein variant.
****
ContraBAR: Contrastive Bayes-Adaptive Deep RL
Era Choshen, Aviv Tamar
In meta reinforcement learning (meta RL), an agent seeks a Bayes-optimal policy – the optimal policy when facing an unknown task that is sampled from some known task distribution. Previous approaches tackled this problem by inferring a $\textit{belief}$ over task parameters, using variational inference methods. Motivated by recent successes of contrastive learning approaches in RL, such as contrastive predictive coding (CPC), we investigate whether contrastive methods can be used for learning Bayes-optimal behavior. We begin by proving that representations learned by CPC are indeed sufficient for Bayes optimality. Based on this observation, we propose a simple meta RL algorithm that uses CPC in lieu of variational belief inference. Our method, $\textit{ContraBAR}$, achieves comparable performance to state-of-the-art in domains with state-based observation and circumvents the computational toll of future observation reconstruction, enabling learning in domains with image-based observations. It can also be combined with image augmentations for domain randomization and used seamlessly in both online and offline meta RL settings.
****
Forget Unlearning: Towards True Data-Deletion in Machine Learning
Rishav Chourasia, Neil Shah
Unlearning algorithms aim to remove deleted data’s influence from trained models at a cost lower than full retraining. However, prior guarantees of unlearning in literature are flawed and don’t protect the privacy of deleted records. We show that when people delete their data as a function of published models, records in a database become interdependent. So, even retraining a fresh model after deletion of a record doesn’t ensure its privacy. Secondly, unlearning algorithms that cache partial computations to speed up the processing can leak deleted information over a series of releases, violating the privacy of deleted records in the long run. To address these, we propose a sound deletion guarantee and show that ensuring the privacy of existing records is necessary for the privacy of deleted records. Under this notion, we propose an optimal, computationally efficient, and sound machine unlearning algorithm based on noisy gradient descent.
****
Patch-level Routing in Mixture-of-Experts is Provably Sample-efficient for Convo

lutional Neural Networks
Mohammed Nowaz Rabbani Chowdhury, Shuai Zhang, Meng Wang, Sijia Liu, Pin-Yu Chen
In deep learning, mixture-of-experts (MoE) activates one or few experts (sub-networks) on a per-sample or per-token basis, resulting in significant computation reduction. The recently proposed patch-level routing in MoE (pMoE) divides each input into n patches (or tokens) and sends l patches ($l\ll n$) to each expert through prioritized routing. pMoE has demonstrated great empirical success in reducing training and inference costs while maintaining test accuracy. However, the theoretical explanation of pMoE and the general MoE remains elusive. Focusing on a supervised classification task using a mixture of two-layer convolutional neural networks (CNNs), we show for the first time that pMoE provably reduces the required number of training samples to achieve desirable generalization (referred to as the sample complexity) by a factor in the polynomial order of n/l, and outperforms its single-expert counterpart of the same or even larger capacity. The advantage results from the discriminative routing property, which is justified in both theory and practice that pMoE routers can filter label-irrelevant patches and route similar class-discriminative patches to the same expert. Our experimental results on MNIST, CIFAR-10, and CelebA support our theoretical findings on pMoE's generalization and show that pMoE can avoid learning spurious correlations.
****
What do CNNs Learn in the First Layer and Why? A Linear Systems Perspective
Rhea Chowers, Yair Weiss
It has previously been reported that the representation that is learned in the first layer of deep Convolutional Neural Networks (CNNs) is highly consistent across initializations and architectures. In this work, we quantify this consistency by considering the first layer as a filter bank and measuring its energy distribution. We find that the energy distribution is very different from that of the initial weights and is remarkably consistent across random initializations, datasets, architectures and even when the CNNs are trained with random labels. In order to explain this consistency, we derive an analytical formula for the energy profile of linear CNNs and show that this profile is mostly dictated by the second order statistics of image patches in the training set and it will approach a whitening transformation when the number of iterations goes to infinity. Finally, we show that this formula for linear CNNs also gives an excellent fit for the energy profiles learned by commonly used nonlinear CNNs such as ResNet and VGG, and that the first layer of these CNNs indeed performs approximate whitening of their inputs.
****
Unifying Molecular and Textual Representations via Multi-task Language Modelling
Dimitrios Christofidellis, Giorgio Giannone, Jannis Born, Ole Winther, Teodoro Laino, Matteo Manica
The recent advances in neural language models have also been successfully applied to the field of chemistry, offering generative solutions for classical problems in molecular design and synthesis planning. These new methods have the potential to fuel a new era of data-driven automation in scientific discovery. However, specialized models are still typically required for each task, leading to the need for problem-specific fine-tuning and neglecting task interrelations. The main obstacle in this field is the lack of a unified representation between natural language and chemical representations, complicating and limiting human-machine interaction. Here, we propose the first multi-domain, multi-task language model that can solve a wide range of tasks in both the chemical and natural language domains. Our model can handle chemical and natural language concurrently, without requiring expensive pre-training on single domains or task-specific models. Interestingly, sharing weights across domains remarkably improves our model when benchmarked against state-of-the-art baselines on single-domain and cross-domain tasks. In particular, sharing information across domains and tasks gives rise to large improvements in cross-domain tasks, the magnitude of which increase with scale, as measured by more than a dozen of relevant metrics. Our work suggests that such models can robustly and efficiently accelerate discovery in physical sci

ences by superseding problem-specific fine-tuning and enhancing human-model interactions.
****
Wasserstein Barycenter Matching for Graph Size Generalization of Message Passing Neural Networks
Xu Chu, Yujie Jin, Xin Wang, Shanghang Zhang, Yasha Wang, Wenwu Zhu, Hong Mei
Graph size generalization is hard for Message passing neural networks (MPNNs). The graph-level classification performance of MPNNs degrades across various graph sizes. Recently, theoretical studies reveal that a slow uncontrollable convergence rate w.r.t. graph size could adversely affect the size generalization. To address the uncontrollable convergence rate caused by correlations across nodes in the underlying dimensional signal-generating space, we propose to use Wasserstein barycenters as graph-level consensus to combat node-level correlations. Methodologically, we propose a Wasserstein barycenter matching (WBM) layer that represents an input graph by Wasserstein distances between its MPNN-filtered node embeddings versus some learned class-wise barycenters. Theoretically, we show that the convergence rate of an MPNN with a WBM layer is controllable and independent to the dimensionality of the signal-generating space. Thus MPNNs with WBM layers are less susceptible to slow uncontrollable convergence rate and size variations. Empirically, the WBM layer improves the size generalization over vanilla MPNNs with different backbones (e.g., GCN, GIN, and PNA) significantly on real-world graph datasets.
****
Shape-Guided Dual-Memory Learning for 3D Anomaly Detection
Yu-Min Chu, Chieh Liu, Ting-I Hsieh, Hwann-Tzong Chen, Tyng-Luh Liu
We present a shape-guided expert-learning framework to tackle the problem of unsupervised 3D anomaly detection. Our method is established on the effectiveness of two specialized expert models and their synergy to localize anomalous regions from color and shape modalities. The first expert utilizes geometric information to probe 3D structural anomalies by modeling the implicit distance fields around local shapes. The second expert considers the 2D RGB features associated with the first expert to identify color appearance irregularities on the local shapes. We use the two experts to build the dual memory banks from the anomaly-free training samples and perform shape-guided inference to pinpoint the defects in the testing samples. Owing to the per-point 3D representation and the effective fusion scheme of complementary modalities, our method efficiently achieves state-of-the-art performance on the MVTec 3D-AD dataset with better recall and lower false positive rates, as preferred in real applications.
****
Multiply Robust Off-policy Evaluation and Learning under Truncation by Death
Jianing Chu, Shu Yang, Wenbin Lu
Typical off-policy evaluation (OPE) and off-policy learning (OPL) are not well-defined problems under "truncation by death", where the outcome of interest is not defined after some events, such as death. The standard OPE no longer yields consistent estimators, and the standard OPL results in suboptimal policies. In this paper, we formulate OPE and OPL using principal stratification under "truncation by death". We propose a survivor value function for a subpopulation whose outcomes are always defined regardless of treatment conditions. We establish a novel identification strategy under principal ignorability, and derive the semiparametric efficiency bound of an OPE estimator. Then, we propose multiply robust estimators for OPE and OPL. We show that the proposed estimators are consistent and asymptotically normal even with flexible semi/nonparametric models for nuisance functions approximation. Moreover, under mild rate conditions of nuisance functions approximation, the estimators achieve the semiparametric efficiency bound. Finally, we conduct experiments to demonstrate the empirical performance of the proposed estimators.
****
InfoOT: Information Maximizing Optimal Transport
Ching-Yao Chuang, Stefanie Jegelka, David Alvarez-Melis
Optimal transport aligns samples across distributions by minimizing the transpor

tation cost between them, e.g., the geometric distances. Yet, it ignores coheren ce structure in the data such as clusters, does not handle outliers well, and ca nnot integrate new data points. To address these drawbacks, we propose InfoOT, a n information-theoretic extension of optimal transport that maximizes the mutual information between domains while minimizing geometric distances. The resulting objective can still be formulated as a (generalized) optimal transport problem, and can be efficiently solved by projected gradient descent. This formulation y ields a new projection method that is robust to outliers and generalizes to unse en samples. Empirically, InfoOT improves the quality of alignments across benchm arks in domain adaptation, cross-domain retrieval, and single-cell alignment.
****

A Toy Model of Universality: Reverse Engineering how Networks Learn Group Operat ions

Bilal Chughtai, Lawrence Chan, Neel Nanda

Universality is a key hypothesis in mechanistic interpretability – that differen t models learn similar features and circuits when trained on similar tasks. In t his work, we study the universality hypothesis by examining how small networks l earn to implement group compositions. We present a novel algorithm by which neur al networks may implement composition for any finite group via mathematical repr esentation theory. We then show that these networks consistently learn this algo rithm by reverse engineering model logits and weights, and confirm our understan ding using ablations. By studying networks trained on various groups and archite ctures, we find mixed evidence for universality: using our algorithm, we can com pletely characterize the family of circuits and features that networks learn on this task, but for a given network the precise circuits learned – as well as the order they develop – are arbitrary.
****

Distribution Free Prediction Sets for Node Classification

Jase Clarkson

Graph Neural Networks (GNNs) are able to achieve high classification accuracy on many important real world datasets, but provide no rigorous notion of predictiv e uncertainty. Quantifying the confidence of GNN models is difficult due to the dependence between datapoints induced by the graph structure. We leverage recent advances in conformal prediction to construct prediction sets for node classifi cation in inductive learning scenarios. We do this by taking an existing approac h for conformal classification that relies on exchangeable data and modifying it by appropriately weighting the conformal scores to reflect the network structur e. We show through experiments on standard benchmark datasets using popular GNN models that our approach provides tighter and better calibrated prediction sets than a naive application of conformal prediction.
****

Sequential Strategic Screening

Lee Cohen, Saeed Sharifi -Malvajerdi, Kevin Stangl, Ali Vakilian, Juba Ziani

We initiate the study of strategic behavior in screening processes with multiple classifiers. We focus on two contrasting settings: a "conjunctive” setting in w hich an individual must satisfy all classifiers simultaneously, and a sequential setting in which an individual to succeed must satisfy classifiers one at a tim e. In other words, we introduce the combination of strategic classificationwith screening processes. We show that sequential screening pipelines exhibit new and surprising behavior where individuals can exploit the sequential ordering of th e tests to "zig-zag” between classifiers without having to simultaneously satisf y all of them. We demonstrate an individual can obtain a positive outcome using a limited manipulation budget even when far from the intersection of the positiv e regions of every classifier. Finally, we consider a learner whose goal is to d esign a sequential screening process that is robust to such manipulations, and p rovide a construction for the learner that optimizes a natural objective.
****

Few-Sample Feature Selection via Feature Manifold Learning

David Cohen, Tal Shnitzer, Yuval Kluger, Ronen Talmon

In this paper, we present a new method for few-sample supervised feature selecti

on (FS). Our method first learns the manifold of the feature space of each class using kernels capturing multi-feature associations. Then, based on Riemannian geometry, a composite kernel is computed, extracting the differences between the learned feature associations. Finally, a FS score based on spectral analysis is proposed. Considering multi-feature associations makes our method multivariate by design. This in turn allows for the extraction of the hidden manifold underlying the features and avoids overfitting, facilitating few-sample FS. We showcase the efficacy of our method on illustrative examples and several benchmarks, where our method demonstrates higher accuracy in selecting the informative features compared to competing methods. In addition, we show that our FS leads to improved classification and better generalization when applied to test data.
****

Spatial Implicit Neural Representations for Global-Scale Species Mapping
Elijah Cole, Grant Van Horn, Christian Lange, Alexander Shepard, Patrick Leary, Pietro Perona, Scott Loarie, Oisin Mac Aodha
Estimating the geographical range of a species from sparse observations is a challenging and important geospatial prediction problem. Given a set of locations where a species has been observed, the goal is to build a model to predict whether the species is present or absent at any location. This problem has a long history in ecology, but traditional methods struggle to take advantage of emerging large-scale crowdsourced datasets which can include tens of millions of records for hundreds of thousands of species. In this work, we use Spatial Implicit Neural Representations (SINRs) to jointly estimate the geographical range of 47k species simultaneously. We find that our approach scales gracefully, making increasingly better predictions as we increase the number of species and the amount of data per species when training. To make this problem accessible to machine learning researchers, we provide four new benchmarks that measure different aspects of species range estimation and spatial representation learning. Using these benchmarks, we demonstrate that noisy and biased crowdsourced data can be combined with implicit neural representations to approximate expert-developed range maps for many species.
****

K-SHAP: Policy Clustering Algorithm for Anonymous Multi-Agent State-Action Pairs
Andrea Coletta, Svitlana Vyetrenko, Tucker Balch
Learning agent behaviors from observational data has shown to improve our understanding of their decision-making processes, advancing our ability to explain their interactions with the environment and other agents. While multiple learning techniques have been proposed in the literature, there is one particular setting that has not been explored yet: multi agent systems where agent identities remain anonymous. For instance, in financial markets labeled data that identifies market participant strategies is typically proprietary, and only the anonymous state-action pairs that result from the interaction of multiple market participants are publicly available. As a result, sequences of agent actions are not observable, restricting the applicability of existing work. In this paper, we propose a Policy Clustering algorithm, called K-SHAP, that learns to group anonymous state-action pairs according to the agent policies. We frame the problem as an Imitation Learning (IL) task, and we learn a world-policy able to mimic all the agent behaviors upon different environmental states. We leverage the world-policy to explain each anonymous observation through an additive feature attribution method called SHAP (SHapley Additive exPlanations). Finally, by clustering the explanations we show that we are able to identify different agent policies and group observations accordingly. We evaluate our approach on simulated synthetic market data and a real-world financial dataset. We show that our proposal significantly and consistently outperforms the existing methods, identifying different agent strategies.
****

Inferring Relational Potentials in Interacting Systems
Armand Comas, Yilun Du, Christian Fernandez Lopez, Sandesh Ghimire, Mario Sznaier, Joshua B. Tenenbaum, Octavia Camps
Systems consisting of interacting agents are prevalent in the world, ranging fro

m dynamical systems in physics to complex biological networks. To build systems which can interact robustly in the real world, it is thus important to be able t o infer the precise interactions governing such systems. Existing approaches typ ically discover such interactions by explicitly modeling the feed-forward dynami cs of the trajectories. In this work, we propose Neural Interaction Inference wi th Potentials (NIIP) as an alternative approach to discover such interactions th at enables greater flexibility in trajectory modeling: it discovers a set of rel ational potentials, represented as energy functions, which when minimized recons truct the original trajectory. NIIP assigns low energy to the subset of trajecto ries which respect the relational constraints observed. We illustrate that with these representations NIIP displays unique capabilities in test-time. First, it allows trajectory manipulation, such as interchanging interaction types across s eparately trained models, as well as trajectory forecasting. Additionally, it al lows adding external hand-crafted potentials at test-time. Finally, NIIP enables the detection of out-of-distribution samples and anomalies without explicit tra ining.

****

Task-specific experimental design for treatment effect estimation

Bethany Connolly, Kim Moore, Tobias Schwedes, Alexander Adam, Gary Willis, Ilya Feige, Christopher Frye

Understanding causality should be a core requirement of any attempt to build rea l impact through AI. Due to the inherent unobservability of counterfactuals, lar ge randomised trials (RCTs) are the standard for causal inference. But large exp eriments are generically expensive, and randomisation carries its own costs, e.g . when suboptimal decisions are trialed. Recent work has proposed more sample-ef ficient alternatives to RCTs, but these are not adaptable to the downstream appl ication for which the causal effect is sought. In this work, we develop a task-s pecific approach to experimental design and derive sampling strategies customise d to particular downstream applications. Across a range of important tasks, real -world datasets, and sample sizes, our method outperforms other benchmarks, e.g. requiring an order-of-magnitude less data to match RCT performance on targeted marketing tasks.

****

A Mathematical Model for Curriculum Learning for Parities

Elisabetta Cornacchia, Elchanan Mossel

Curriculum learning (CL)- training using samples that are generated and presente d in a meaningful order - was introduced in the machine learning context around a decade ago. While CL has been extensively used and analysed empirically, there has been very little mathematical justification for its advantages. We introduc e a CL model for learning the class of k-parities on d bits of a binary string w ith a neural network trained by stochastic gradient descent (SGD). We show that a wise choice of training examples, involving two or more product distributions, allows to reduce significantly the computational cost of learning this class of functions, compared to learning under the uniform distribution. We conduct expe riments to support our analysis. Furthermore, we show that for another class of functions - namely the ‘Hamming mixtures’ - CL strategies involving a bounded nu mber of product distributions are not beneficial.

****

Learning to Maximize Mutual Information for Dynamic Feature Selection

Ian Connick Covert, Wei Qiu, Mingyu Lu, Na Yoon Kim, Nathan J White, Su-In Lee

Feature selection helps reduce data acquisition costs in ML, but the standard ap proach is to train models with static feature subsets. Here, we consider the dyn amic feature selection (DFS) problem where a model sequentially queries features based on the presently available information. DFS is often addressed with reinf orcement learning, but we explore a simpler approach of greedily selecting featu res based on their conditional mutual information. This method is theoretically appealing but requires oracle access to the data distribution, so we develop a l earning approach based on amortized optimization. The proposed method is shown t o recover the greedy policy when trained to optimality, and it outperforms numer ous existing feature selection methods in our experiments, thus validating it as

a simple but powerful approach for this problem.
****
Rethinking Weak Supervision in Helping Contrastive Learning
Jingyi Cui, Weiran Huang, Yifei Wang, Yisen Wang
Contrastive learning has shown outstanding performances in both supervised and unsupervised learning, and has recently been introduced to solve weakly supervised learning problems such as semi-supervised learning and noisy label learning. Despite the empirical evidence showing that semi-supervised labels improve the representations of contrastive learning, it remains unknown if noisy supervised information can be directly used in training instead of after manual denoising. Therefore, to explore the mechanical differences between semi-supervised and noisy-labeled information in helping contrastive learning, we establish a unified theoretical framework of contrastive learning under weak supervision. Specifically, we investigate the most intuitive paradigm of jointly training supervised and unsupervised contrastive losses. By translating the weakly supervised information into a similarity graph under the framework of spectral clustering based on the posterior probability of weak labels, we establish the downstream classification error bound. We prove that semi-supervised labels improve the downstream error bound whereas noisy labels have limited effects under such a paradigm. Our theoretical findings here provide new insights for the community to rethink the role of weak supervision in helping contrastive learning.
****
Bayes-optimal Learning of Deep Random Networks of Extensive-width
Hugo Cui, Florent Krzakala, Lenka Zdeborova
We consider the problem of learning a target function corresponding to a deep, extensive-width, non-linear neural network with random Gaussian weights. We consider the asymptotic limit where the number of samples, the input dimension and the network width are proportionally large and propose a closed-form expression for the Bayes-optimal test error, for regression and classification tasks. We further compute closed-form expressions for the test errors of ridge regression, kernel and random features regression. We find, in particular, that optimally regularized ridge regression, as well as kernel regression, achieve Bayes-optimal performances, while the logistic loss yields a near-optimal test error for classification. We further show numerically that when the number of samples grows faster than the dimension, ridge and kernel methods become suboptimal, while neural networks achieve test error close to zero from quadratically many samples.
****
A General Representation Learning Framework with Generalization Performance Guarantees
Junbiao Cui, Jianqing Liang, Qin Yue, Jiye Liang
The generalization performance of machine learning methods depends heavily on the quality of data representation. However, existing researches rarely consider representation learning from the perspective of generalization error. In this paper, we prove that generalization error of representation learning function can be estimated effectively by solving two convex optimization problems. Based on it, we propose a general representation learning framework. And then, we apply the proposed framework to two most commonly used nonlinear mapping methods, i.e., kernel based method and deep neural network (DNN), and thus design a kernel selection method and a DNN boosting framework, correspondingly. Finally, extensive experiments verify the effectiveness of the proposed methods.
****
IRNeXt: Rethinking Convolutional Network Design for Image Restoration
Yuning Cui, Wenqi Ren, Sining Yang, Xiaochun Cao, Alois Knoll
We present IRNeXt, a simple yet effective convolutional network architecture for image restoration. Recently, Transformer models have dominated the field of image restoration due to the powerful ability of modeling long-range pixels interactions. In this paper, we excavate the potential of the convolutional neural network (CNN) and show that our CNN-based model can receive comparable or better performance than Transformer models with low computation overhead on several image restoration tasks. By re-examining the characteristics possessed by advanced ima

ge restoration algorithms, we discover several key factors leading to the performance improvement of restoration models. This motivates us to develop a novel network for image restoration based on cheap convolution operators. Comprehensive experiments demonstrate that IRNeXt delivers state-of-the-art performance among numerous datasets on a range of image restoration tasks with low computational complexity, including image dehazing, single-image defocus/motion deblurring, image deraining, and image desnowing. https://github.com/c-yn/IRNeXt.
****

Scaling Up Dataset Distillation to ImageNet-1K with Constant Memory
Justin Cui, Ruochen Wang, Si Si, Cho-Jui Hsieh
Dataset Distillation is a newly emerging area that aims to distill large datasets into much smaller and highly informative synthetic ones to accelerate training and reduce storage. Among various dataset distillation methods, trajectory-matching-based methods (MTT) have achieved SOTA performance in many tasks, e.g., on CIFAR-10/100. However, due to exorbitant memory consumption when unrolling optimization through SGD steps, MTT fails to scale to large-scale datasets such as ImageNet-1K. Can we scale this SOTA method to ImageNet-1K and does its effectiveness on CIFAR transfer to ImageNet-1K? To answer these questions, we first propose a procedure to exactly compute the unrolled gradient with constant memory complexity, which allows us to scale MTT to ImageNet-1K seamlessly with ~ 6x reduction in memory footprint. We further discover that it is challenging for MTT to handle datasets with a large number of classes, and propose a novel soft label assignment that drastically improves its convergence. The resulting algorithm sets new SOTA on ImageNet-1K: we can scale up to 50 IPCs (Image Per Class) on ImageNet-1K on a single GPU (all previous methods can only scale to 2 IPCs on ImageNet-1K), leading to the best accuracy (only 5.9% accuracy drop against full dataset training) while utilizing only 4.2% of the number of data points - an 18.2% absolute gain over prior SOTA.
****

Learning Dynamic Query Combinations for Transformer-based Object Detection and Segmentation
Yiming Cui, Linjie Yang, Haichao Yu
Transformer-based detection and segmentation methods use a list of learned detection queries to retrieve information from the transformer network and learn to predict the location and category of one specific object from each query. We empirically find that random convex combinations of the learned queries are still good for the corresponding models. We then propose to learn a convex combination with dynamic coefficients based on the high-level semantics of the image. The generated dynamic queries, named as modulated queries, better capture the prior of object locations and categories in the different images. Equipped with our modulated queries, a wide range of DETR-based models achieve consistent and superior performance across multiple tasks (object detection, instance segmentation, panoptic segmentation) and on different benchmarks (MS COCO, CityScapes, YoutubeVIS).
****

Adaptive Identification of Populations with Treatment Benefit in Clinical Trials: Machine Learning Challenges and Solutions
Alicia Curth, Alihan Hüyük, Mihaela Van Der Schaar
We study the problem of adaptively identifying patient subpopulations that benefit from a given treatment during a confirmatory clinical trial. This type of adaptive clinical trial has been thoroughly studied in biostatistics, but has been allowed only limited adaptivity so far. Here, we aim to relax classical restrictions on such designs and investigate how to incorporate ideas from the recent machine learning literature on adaptive and online experimentation to make trials more flexible and efficient. We find that the unique characteristics of the subpopulation selection problem – most importantly that (i) one is usually interested in finding subpopulations with any treatment benefit (and not necessarily the single subgroup with largest effect) given a limited budget and that (ii) effectiveness only has to be demonstrated across the subpopulation on average – give rise to interesting challenges and new desiderata when designing algorithmic solu

tions. Building on these findings, we propose AdaGGI and AdaGCPI, two meta-algorithms for subpopulation construction. We empirically investigate their performance across a range of simulation scenarios and derive insights into their (dis)advantages across different settings.
****

In Search of Insights, Not Magic Bullets: Towards Demystification of the Model Selection Dilemma in Heterogeneous Treatment Effect Estimation

Alicia Curth, Mihaela Van Der Schaar

Personalized treatment effect estimates are often of interest in high-stakes applications – thus, before deploying a model estimating such effects in practice, one needs to be sure that the best candidate from the ever-growing machine learning toolbox for this task was chosen. Unfortunately, due to the absence of counterfactual information in practice, it is usually not possible to rely on standard validation metrics for doing so, leading to a well-known model selection dilemma in the treatment effect estimation literature. While some solutions have recently been investigated, systematic understanding of the strengths and weaknesses of different model selection criteria is still lacking. In this paper, instead of attempting to declare a global 'winner', we therefore empirically investigate success- and failure modes of different selection criteria. We highlight that there is a complex interplay between selection strategies, candidate estimators and the data used for comparing them, and provide interesting insights into the relative (dis)advantages of different criteria alongside desiderata for the design of further illuminating empirical studies in this context.
****

Optimal Stochastic Non-smooth Non-convex Optimization through Online-to-Non-convex Conversion

Ashok Cutkosky, Harsh Mehta, Francesco Orabona

We present new algorithms for optimizing non-smooth, non-convex stochastic objectives based on a novel analysis technique. This improves the current best-known complexity for finding a (δ,ϵ)-stationary point from $O(\epsilon^{-4}\delta^{-1})$ stochastic gradient queries to $O(\epsilon^{-3}\delta^{-1})$, which we also show to be optimal. Our primary technique is a reduction from non-smooth non-convex optimization to online learning, after which our results follow from standard regret bounds in online learning. For deterministic and second-order smooth objectives, applying more advanced optimistic online learning techniques enables a new complexity of $O(\epsilon^{-1.5}\delta^{-0.5})$. Our improved non-smooth analysis also immediately recovers all optimal or best-known results for finding ϵ stationary points of smooth or second-order smooth objectives in both stochastic and deterministic settings.
****

Monge, Bregman and Occam: Interpretable Optimal Transport in High-Dimensions with Feature-Sparse Maps

Marco Cuturi, Michal Klein, Pierre Ablin

Optimal transport (OT) theory focuses, among all maps $T:\mathbb{R}^d\rightarrow \mathbb{R}^d$ that can morph a probability measure μ onto another ν, on those that are the "thriftiest", i.e. such that the average cost $c(x, T(x))$ between x and its image $T(x)$ is as small as possible. Many computational approaches have been proposed to estimate such Monge maps when c is the squared-Euclidean distance, e.g., using entropic maps [Pooladian+2021], or input convex neural networks [Makkuva+2020, Korotin+2020]. We propose a new research direction, that leverages a specific translation invariant cost $c(x, y):=h(x-y)$ inspired by the elastic net. Here, $h:=\tfrac{1}{2}\|\cdot\|_2^2+\tau(\cdot)$, where τ is a convex function. We highlight a surprising link tying together a generalized entropic map for h, Bregman centroids induced by h, and the proximal operator of τ. We show how setting τ to be a sparsity-inducing norm results in the first application of Occam's razor to transport. These maps yield, mechanically, displacement vectors $\Delta(x):= T(x)-x$ that are sparse, with sparsity patterns that vary depending on x. We showcase the ability of our method to estimate meaningful OT maps for high-dimensional single-cell transcription data. We use our methods in the 34000-d space of gene counts for cells, without

using a prior dimensionality reduction, thus retaining the ability to interpret all displacements at the gene level.
****

From Noisy Fixed-Point Iterations to Private ADMM for Centralized and Federated Learning
Edwige Cyffers, Aurélien Bellet, Debabrota Basu
We study differentially private (DP) machine learning algorithms as instances of noisy fixed-point iterations, in order to derive privacy and utility results from this well-studied framework. We show that this new perspective recovers popular private gradient-based methods like DP-SGD and provides a principled way to design and analyze new private optimization algorithms in a flexible manner. Focusing on the widely-used Alternating Directions Method of Multipliers (ADMM) method, we use our general framework derive novel private ADMM algorithms for centralized, federated and fully decentralized learning. We establish strong privacy guarantees for these algorithms, leveraging privacy amplification by iteration and by subsampling. Finally, we provide utility guarantees for the three algorithms using a unified analysis that exploits a recent linear convergence result for noisy fixed-point iterations.
****

Chameleon: Adapting to Peer Images for Planting Durable Backdoors in Federated Learning
Yanbo Dai, Songze Li
In a federated learning (FL) system, distributed clients upload their local models to a central server to aggregate into a global model. Malicious clients may plant backdoors into the global model through uploading poisoned local models, causing images with specific patterns to be misclassified into some target labels. Backdoors planted by current attacks are not durable, and vanish quickly once the attackers stop model poisoning. In this paper, we investigate the connection between the durability of FL backdoors and the relationships between benign images and poisoned images (i.e., the images whose labels are flipped to the target label during local training). Specifically, benign images with the original and the target labels of the poisoned images are found to have key effects on backdoor durability. Consequently, we propose a novel attack, Chameleon, which utilizes contrastive learning to further amplify such effects towards a more durable backdoor. Extensive experiments demonstrate that Chameleon significantly extends the backdoor lifespan over baselines by $1.2\times \sim 4\times$, for a wide range of image datasets, backdoor types, and model architectures.
****

Refined Regret for Adversarial MDPs with Linear Function Approximation
Yan Dai, Haipeng Luo, Chen-Yu Wei, Julian Zimmert
We consider learning in an adversarial Markov Decision Process (MDP) where the loss functions can change arbitrarily over K episodes and the state space can be arbitrarily large. We assume that the Q-function of any policy is linear in some known features, that is, a linear function approximation exists. The best existing regret upper bound for this setting (Luo et al., 2021) is of order $\tilde{\mathcal O}(K^{2/3})$ (omitting all other dependencies), given access to a simulator. This paper provides two algorithms that improve the regret to $\tilde{\mathcal O}(\sqrt K)$ in the same setting. Our first algorithm makes use of a refined analysis of the Follow-the-Regularized-Leader (FTRL) algorithm with the log-barrier regularizer. This analysis allows the loss estimators to be arbitrarily negative and might be of independent interest. Our second algorithm develops a magnitude-reduced loss estimator, further removing the polynomial dependency on the number of actions in the first algorithm and leading to the optimal regret bound (up to logarithmic terms and dependency on the horizon). Moreover, we also extend the first algorithm to simulator-free linear MDPs, which achieves $\tilde{\mathcal O}(K^{8/9})$ regret and greatly improves over the best existing bound $\tilde{\mathcal O}(K^{14/15})$. This algorithm relies on a better alternative to the Matrix Geometric Resampling procedure by Neu & Olkhovskaya (2020), which could again be of independent interest.
****

MultiRobustBench: Benchmarking Robustness Against Multiple Attacks
Sihui Dai, Saeed Mahloujifar, Chong Xiang, Vikash Sehwag, Pin-Yu Chen, Prateek M ittal
The bulk of existing research in defending against adversarial examples focuses on defending against a single (typically bounded ℓ_p-norm) attack, but for a practical setting, machine learning (ML) models should be robust to a wide var iety of attacks. In this paper, we present the first unified framework for consi dering multiple attacks against ML models. Our framework is able to model differ ent levels of learner’s knowledge about the test-time adversary, allowing us to model robustness against unforeseen attacks and robustness against unions of att acks. Using our framework, we present the first leaderboard, MultiRobustBench (h ttps://multirobustbench.github.io), for benchmarking multiattack evaluation whic h captures performance across attack types and attack strengths. We evaluate the performance of 16 defended models for robustness against a set of 9 different a ttack types, including ℓ_p-based threat models, spatial transformations, an d color changes, at 20 different attack strengths (180 attacks total). Additiona lly, we analyze the state of current defenses against multiple attacks. Our anal ysis shows that while existing defenses have made progress in terms of average r obustness across the set of attacks used, robustness against the worst-case atta ck is still a big open problem as all existing models perform worse than random guessing.
****
Moderately Distributional Exploration for Domain Generalization
Rui Dai, Yonggang Zhang, Zhen Fang, Bo Han, Xinmei Tian
Domain generalization (DG) aims to tackle the distribution shift between trainin g domains and unknown target domains. Generating new domains is one of the most effective approaches, yet its performance gain depends on the distribution discr epancy between the generated and target domains. Distributionally robust optimiz ation is promising to tackle distribution discrepancy by exploring domains in an uncertainty set. However, the uncertainty set may be overwhelmingly large, lead ing to low-confidence prediction in DG. It is because a large uncertainty set co uld introduce domains containing semantically different factors from training do mains. To address this issue, we propose to perform a $\textit{mo}$derately $\te xtit{d}$istributional $\textit{e}$xploration (MODE) for domain generalization. S pecifically, MODE performs distribution exploration in an uncertainty $\textit{s ubset}$ that shares the same semantic factors with the training domains. We show that MODE can endow models with provable generalization performance on unknown target domains. The experimental results show that MODE achieves competitive per formance compared to state-of-the-art baselines.
****
Trajectory-Aware Eligibility Traces for Off-Policy Reinforcement Learning
Brett Daley, Martha White, Christopher Amato, Marlos C. Machado
Off-policy learning from multistep returns is crucial for sample-efficient reinf orcement learning, but counteracting off-policy bias without exacerbating varian ce is challenging. Classically, off-policy bias is corrected in a per-decision m anner: past temporal-difference errors are re-weighted by the instantaneous Impo rtance Sampling (IS) ratio after each action via eligibility traces. Many off-po licy algorithms rely on this mechanism, along with differing protocols for cutti ng the IS ratios (traces) to combat the variance of the IS estimator. Unfortunat ely, once a trace has been cut, the effect cannot be easily reversed. This has l ed to the development of credit-assignment strategies that account for multiple past experiences at a time. These trajectory-aware methods have not been extensi vely analyzed, and their theoretical justification remains uncertain. In this pa per, we propose a multistep operator that unifies per-decision and trajectory-aw are methods. We prove convergence conditions for our operator in the tabular set ting, establishing the first guarantees for several existing methods as well as many new ones. Finally, we introduce Recency-Bounded Importance Sampling (RBIS), which leverages trajectory awareness to perform robustly across λ-value s in an off-policy control task.
****

Efficient displacement convex optimization with particle gradient descent
Hadi Daneshmand, Jason D. Lee, Chi Jin
Particle gradient descent, which uses particles to represent a probability measu re and performs gradient descent on particles in parallel, is widely used to opt imize functions of probability measures. This paper considers particle gradient descent with a finite number of particles and establishes its theoretical guaran tees to optimize functions that are displacement convex in measures. Concretely, for Lipschitz displacement convex functions defined on probability over R^d, we prove that $O(1/\epsilon^2)$ particles and $O(d/\epsilon^4)$ iterations are s ufficient to find the ϵ-optimal solutions. We further provide improved complexity bounds for optimizing smooth displacement convex functions. An applic ation of our results proves the conjecture of no optimization-barrier up to perm utation invariance, proposed by Entezari et al. (2022), for specific two-layer n eural networks with two-dimensional inputs uniformly drawn from unit circle.
****

Multiple Thinking Achieving Meta-Ability Decoupling for Object Navigation
Ronghao Dang, Lu Chen, Liuyi Wang, Zongtao He, Chengju Liu, Qijun Chen
We propose a meta-ability decoupling (MAD) paradigm, which brings together vario us object navigation methods in an architecture system, allowing them to mutuall y enhance each other and evolve together. Based on the MAD paradigm, we design a multiple thinking (MT) model that leverages distinct thinking to abstract vario us meta-abilities. Our method decouples meta-abilities from three aspects: input , encoding, and reward while employing the multiple thinking collaboration (MTC) module to promote mutual cooperation between thinking. MAD introduces a novel q ualitative and quantitative interpretability system for object navigation. Throu gh extensive experiments on AI2-Thor and RoboTHOR, we demonstrate that our metho d outperforms state-of-the-art (SOTA) methods on both typical and zero-shot obje ct navigation tasks.
****

Neural Collapse in Deep Linear Networks: From Balanced to Imbalanced Data
Hien Dang, Tho Tran Huu, Stanley Osher, Hung The Tran, Nhat Ho, Tan Minh Nguyen
Modern deep neural networks have achieved impressive performance on tasks from i mage classification to natural language processing. Surprisingly, these complex systems with massive amounts of parameters exhibit the same structural propertie s in their last-layer features and classifiers across canonical datasets when tr aining until convergence. In particular, it has been observed that the last-laye r features collapse to their class-means, and those class-means are the vertices of a simplex Equiangular Tight Frame (ETF). This phenomenon is known as Neural Collapse (NC). Recent papers have theoretically shown that NC emerges in the glo bal minimizers of training problems with the simplified “unconstrained feature m odel”. In this context, we take a step further and prove the NC occurrences in d eep linear networks for the popular mean squared error (MSE) and cross entropy (CE) losses, showing that global solutions exhibit NC properties across the linea r layers. Furthermore, we extend our study to imbalanced data for MSE loss and p resent the first geometric analysis of NC under bias-free setting. Our results d emonstrate the convergence of the last-layer features and classifiers to a geome try consisting of orthogonal vectors, whose lengths depend on the amount of data in their corresponding classes. Finally, we empirically validate our theoretica l analyses on synthetic and practical network architectures with both balanced a nd imbalanced scenarios.
****

Reinforcement Learning Can Be More Efficient with Multiple Rewards
Christoph Dann, Yishay Mansour, Mehryar Mohri
Reward design is one of the most critical and challenging aspects when formulati ng a task as a reinforcement learning (RL) problem. In practice, it often takes several attempts of reward specification and learning with it in order to find o ne that leads to sample-efficient learning of the desired behavior. Instead, in this work, we study whether directly incorporating multiple alternate reward for mulations of the same task in a single agent can lead to faster learning. We ana lyze multi-reward extensions of action-elimination algorithms and prove more fav

orable instance-dependent regret bounds compared to their single-reward counterp arts, both in multi-armed bandits and in tabular Markov decision processes. Our bounds scale for each state-action pair with the inverse of the largest gap amon g all reward functions. This suggests that learning with multiple rewards can in deed be more sample-efficient, as long as the rewards agree on an optimal policy . We further prove that when rewards do not agree, multi-reward action eliminati on in multi-armed bandits still learns a policy that is good across all reward f unctions.
****
Best of Both Worlds Policy Optimization
Christoph Dann, Chen-Yu Wei, Julian Zimmert
Policy optimization methods are popular reinforcement learning algorithms in pra ctice and recent works have build theoretical foundation for them by proving $\s qrt{T}$ regret bounds even when the losses are adversarial. Such bounds are tigh t in the worst case but often overly pessimistic. In this work, we show that by carefully designing the regularizer, bonus terms, and learning rates, one can ac hieve a more favorable $\text{polylog}(T)$ regret bound when the losses are stoc hastic, without sacrificing the worst-case guarantee in the adversarial regime. Specifically, we show the first best of both worlds guarantee for policy optimiz ation in tabular MDPs by leveraging either a Tsallis entropy or a Shannon entrop y regularizer. Then we show that under known transitions, we can further obtain a first-order regret bound in the adversarial regime by leveraging the log barri er regularizer.
****
Image generation with shortest path diffusion
Ayan Das, Stathi Fotiadis, Anil Batra, Farhang Nabiei, Fengting Liao, Sattar Vak ili, Da-Shan Shiu, Alberto Bernacchia
The field of image generation has made significant progress thanks to the introd uction of Diffusion Models, which learn to progressively reverse a given image c orruption. Recently, a few studies introduced alternative ways of corrupting ima ges in Diffusion Models, with an emphasis on blurring. However, these studies ar e purely empirical and it remains unclear what is the optimal procedure for corr upting an image. In this work, we hypothesize that the optimal procedure minimiz es the length of the path taken when corrupting an image towards a given final s tate. We propose the Fisher metric for the path length, measured in the space of probability distributions. We compute the shortest path according to this metri c, and we show that it corresponds to a combination of image sharpening, rather than blurring, and noise deblurring. While the corruption was chosen arbitrarily in previous work, our Shortest Path Diffusion (SPD) determines uniquely the ent ire spatiotemporal structure of the corruption. We show that SPD improves on str ong baselines without any hyperparameter tuning, and outperforms all previous Di ffusion Models based on image blurring. Furthermore, any small deviation from th e shortest path leads to worse performance, suggesting that SPD provides the opt imal procedure to corrupt images. Our work sheds new light on observations made in recent works and provides a new approach to improve diffusion models on image s and other types of data.
****
Efficient List-Decodable Regression using Batches
Abhimanyu Das, Ayush Jain, Weihao Kong, Rajat Sen
We demonstrate the use of batches in studying list-decodable linear regression, in which only $\alpha\in (0,1]$ fraction of batches contain genuine samples from a common distribution and the rest can contain arbitrary or even adversarial sa mples. When genuine batches have $\ge \tilde\Omega(1/\alpha)$ samples each, our algorithm can efficiently find a small list of potential regression parameters, with a high probability that one of them is close to the true parameter. This is the first polynomial time algorithm for list-decodable linear regression, and i ts sample complexity scales nearly linearly with the dimension of the covariates . The polynomial time algorithm is made possible by the batch structure and may not be feasible without it, as suggested by a recent Statistical Query lower bou nd (Diakonikolas et al., 2021b).

****
Beyond Uniform Lipschitz Condition in Differentially Private Optimization
Rudrajit Das, Satyen Kale, Zheng Xu, Tong Zhang, Sujay Sanghavi
Most prior results on differentially private stochastic gradient descent (DP-SGD) are derived under the simplistic assumption of uniform Lipschitzness, i.e., th e per-sample gradients are uniformly bounded. We generalize uniform Lipschitznes s by assuming that the per-sample gradients have sample-dependent upper bounds, i.e., per-sample Lipschitz constants, which themselves may be unbounded. We prov ide principled guidance on choosing the clip norm in DP-SGD for convex over-para meterized settings satisfying our general version of Lipschitzness when the per- sample Lipschitz constants are bounded; specifically, we recommend tuning the cl ip norm only till values up to the minimum per-sample Lipschitz constant. This f inds application in the private training of a softmax layer on top of a deep net work pre-trained on public data. We verify the efficacy of our recommendation vi a experiments on 8 datasets. Furthermore, we provide new convergence results for DP-SGD on convex and nonconvex functions when the Lipschitz constants are unbou nded but have bounded moments, i.e., they are heavy-tailed.
****
Understanding Self-Distillation in the Presence of Label Noise
Rudrajit Das, Sujay Sanghavi
Self-distillation (SD) is the process of first training a "teacher" model and th en using its predictions to train a "student" model that has the same architectu re. Specifically, the student’s loss is $\big(\xi*\ell(\text{teacher’s predictio ns}, \text{ student’s predictions}) + (1-\xi)*\ell(\text{given labels}, \text{ s tudent’s predictions})\big)$, where ℓ is the loss function and ξ is som e parameter $\in [0,1]$. SD has been empirically observed to provide performance gains in several settings. In this paper, we theoretically characterize the eff ect of SD in two supervised learning problems with noisy labels. We first analyz e SD for regularized linear regression and show that in the high label noise reg ime, the optimal value of ξ that minimizes the expected error in estimating the ground truth parameter is surprisingly greater than 1. Empirically, we show that $\xi > 1$ works better than $\xi \leq 1$ even with the cross-entropy loss f or several classification datasets when 50% or 30% of the labels are corrupted. Further, we quantify when optimal SD is better than optimal regularization. Next , we analyze SD in the case of logistic regression for binary classification wit h random label corruption and quantify the range of label corruption in which th e student outperforms the teacher (w.r.t. accuracy). To our knowledge, this is t he first result of its kind for the cross-entropy loss.
****
Interval Bound Interpolation for Few-shot Learning with Few Tasks
Shounak Datta, Sankha Subhra Mullick, Anish Chakrabarty, Swagatam Das
Few-shot learning aims to transfer the knowledge acquired from training on a div erse set of tasks to unseen tasks from the same task distribution, with a limite d amount of labeled data. The underlying requirement for effective few-shot gene ralization is to learn a good representation of the task manifold. This becomes more difficult when only a limited number of tasks are available for training. I n such a few-task few-shot setting, it is beneficial to explicitly preserve the local neighborhoods from the task manifold and exploit this to generate artifici al tasks for training. To this end, we introduce the notion of interval bounds f rom the provably robust training literature to few-shot learning. The interval b ounds are used to characterize neighborhoods around the training tasks. These ne ighborhoods can then be preserved by minimizing the distance between a task and its respective bounds. We then use a novel strategy to artificially form new tas ks for training by interpolating between the available tasks and their respectiv e interval bounds. We apply our framework to both model-agnostic meta-learning a s well as prototype-based metric-learning paradigms. The efficacy of our propose d approach is evident from the improved performance on several datasets from div erse domains in comparison to recent methods.
****
Hypervolume Knowledge Gradient: A Lookahead Approach for Multi-Objective Bayesia

n Optimization with Partial Information
Sam Daulton, Maximilian Balandat, Eytan Bakshy
Bayesian optimization is a popular method for sample efficient multi-objective o
ptimization. However, existing Bayesian optimization techniques fail to effectiv
ely exploit common and often-neglected problem structure such as decoupled evalu
ations, where objectives can be queried independently from one another and each
may consume different resources, or multi-fidelity evaluations, where lower fide
lity-proxies of the objectives can be evaluated at lower cost. In this work, we
propose a general one-step lookahead acquisition function based on the Knowledge
 Gradient that addresses the complex question of what to evaluate when and at wh
ich design points in a principled Bayesian decision-theoretic fashion. Hence, ou
r approach naturally addresses decoupled, multi-fidelity, and standard multi-obj
ective optimization settings in a unified Bayesian decision making framework. By
 construction, our method is the one-step Bayes-optimal policy for hypervolume m
aximization. Empirically, we demonstrate that our method improves sample efficie
ncy in a wide variety of synthetic and real-world problems. Furthermore, we show
 that our method is general-purpose and yields competitive performance in standa
rd (potentially noisy) multi-objective optimization.
****

Fast Combinatorial Algorithms for Min Max Correlation Clustering
Sami Davies, Benjamin Moseley, Heather Newman
We introduce fast algorithms for correlation clustering with respect to the Min
Max objective that provide constant factor approximations on complete graphs. Ou
r algorithms are the first purely combinatorial approximation algorithms for thi
s problem. We construct a novel semi-metric on the set of vertices, which we cal
l the correlation metric, that indicates to our clustering algorithms whether pa
irs of nodes should be in the same cluster. The paper demonstrates empirically t
hat, compared to prior work, our algorithms sacrifice little in the objective qu
ality to obtain significantly better run-time. Moreover, our algorithms scale to
 larger networks that are effectively intractable for known algorithms.
****

Predictive Flows for Faster Ford-Fulkerson
Sami Davies, Benjamin Moseley, Sergei Vassilvitskii, Yuyan Wang
Recent work has shown that leveraging learned predictions can improve the runnin
g time of algorithms for bipartite matching and similar combinatorial problems.
In this work, we build on this idea to improve the performance of the widely use
d Ford-Fulkerson algorithm for computing maximum flows by seeding Ford-Fulkerson
 with predicted flows. Our proposed method offers strong theoretical performance
 in terms of the quality of the prediction. We then consider image segmentation,
 a common use-case of flows in computer vision, and complement our theoretical a
nalysis with strong empirical results.
****

The Persistent Laplacian for Data Science: Evaluating Higher-Order Persistent Sp
ectral Representations of Data
Thomas Davies, Zhengchao Wan, Ruben J Sanchez-Garcia
Persistent homology is arguably the most successful technique in Topological Dat
a Analysis. It combines homology, a topological feature of a data set, with pers
istence, which tracks the evolution of homology over different scales. The persi
stent Laplacian is a recent theoretical development that combines persistence wi
th the combinatorial Laplacian, the higher-order extension of the well-known gra
ph Laplacian. Crucially, the Laplacian encode both the homology of a data set, a
nd some additional geometric information not captured by the homology. Here, we
provide the first investigation into the efficacy of the persistence Laplacian a
s an embedding of data for downstream classification and regression tasks. We ex
tend the persistent Laplacian to cubical complexes so it can be used on images,
then evaluate its performance as an embedding method on the MNIST and MoleculeNe
t datasets, demonstrating that it consistently outperforms persistent homology a
cross tasks.
****

Mitigating Propagation Failures in Physics-informed Neural Networks using Retain

-Resample-Release (R3) Sampling
Arka Daw, Jie Bu, Sifan Wang, Paris Perdikaris, Anuj Karpatne
Despite the success of physics-informed neural networks (PINNs) in approximating partial differential equations (PDEs), PINNs can sometimes fail to converge to the correct solution in problems involving complicated PDEs. This is reflected in several recent studies on characterizing the "failure modes" of PINNs, although a thorough understanding of the connection between PINN failure modes and sampling strategies is missing. In this paper, we provide a novel perspective of failure modes of PINNs by hypothesizing that training PINNs relies on successful "propagation" of solution from initial and/or boundary condition points to interior points. We show that PINNs with poor sampling strategies can get stuck at trivial solutions if there are propagation failures, characterized by highly imbalanced PDE residual fields. To mitigate propagation failures, we propose a novel Retain-Resample-Release sampling (R3) algorithm that can incrementally accumulate collocation points in regions of high PDE residuals with little to no computational overhead. We provide an extension of R3 sampling to respect the principle of causality while solving time-dependent PDEs. We theoretically analyze the behavior of R3 sampling and empirically demonstrate its efficacy and efficiency in comparison with baselines on a variety of PDE problems.
****
On the Robustness of Randomized Ensembles to Adversarial Perturbations
Hassan Dbouk, Naresh Shanbhag
Randomized ensemble classifiers (RECs), where one classifier is randomly selected during inference, have emerged as an attractive alternative to traditional ensembling methods for realizing adversarially robust classifiers with limited compute requirements. However, recent works have shown that existing methods for constructing RECs are more vulnerable than initially claimed, casting major doubts on their efficacy and prompting fundamental questions such as: "When are RECs useful?", "What are their limits?", and "How do we train them?". In this work, we first demystify RECs as we derive fundamental results regarding their theoretical limits, necessary and sufficient conditions for them to be useful, and more. Leveraging this new understanding, we propose a new boosting algorithm (BARRE) for training robust RECs, and empirically demonstrate its effectiveness at defending against strong ℓ_∞ norm-bounded adversaries across various network architectures and datasets. Our code can be found at https://github.com/hsndbk4/BARRE.
****
Pre-computed memory or on-the-fly encoding? A hybrid approach to retrieval augmentation makes the most of your compute
Michiel De Jong, Yury Zemlyanskiy, Nicholas Fitzgerald, Joshua Ainslie, Sumit Sanghai, Fei Sha, William W. Cohen
Retrieval-augmented language models such as Fusion-in-Decoder are powerful, setting the state of the art on a variety of knowledge-intensive tasks. However, they are also expensive, due to the need to encode a large number of retrieved passages. Some work avoids this cost by pre-encoding a text corpus into a memory and retrieving dense representations directly. However, pre-encoding memory incurs a severe quality penalty as the memory representations are not conditioned on the current input. We propose LUMEN, a hybrid between these two extremes, pre-computing the majority of the retrieval representation and completing the encoding on the fly using a live encoder that is conditioned on the question and fine-tuned for the task. We show that LUMEN significantly outperforms pure memory on multiple question-answering tasks while being much cheaper than FiD, and outperforms both for any given compute budget. Moreover, the advantage of LUMEN over FiD increases with model size.
****
Continuous Spatiotemporal Transformer
Antonio Henrique De Oliveira Fonseca, Emanuele Zappala, Josue Ortega Caro, David Van Dijk
Modeling spatiotemporal dynamical systems is a fundamental challenge in machine learning. Transformer models have been very successful in NLP and computer visio

n where they provide interpretable representations of data. However, a limitatio n of transformers in modeling continuous dynamical systems is that they are fund amentally discrete time and space models and thus have no guarantees regarding c ontinuous sampling. To address this challenge, we present the Continuous Spatiot emporal Transformer (CST), a new transformer architecture that is designed for m odeling of continuous systems. This new framework guarantees a continuous and sm ooth output via optimization in Sobolev space. We benchmark CST against traditio nal transformers as well as other spatiotemporal dynamics modeling methods and a chieve superior performance in a number of tasks on synthetic and real systems, including learning brain dynamics from calcium imaging data.

****

The Value of Out-of-Distribution Data

Ashwin De Silva, Rahul Ramesh, Carey Priebe, Pratik Chaudhari, Joshua T Vogelste in

Generalization error always improves with more in-distribution data. However, it is an open question what happens as we add out-of-distribution (OOD) data. Intu itively, if the OOD data is quite different, it seems more data would harm gener alization error, though if the OOD data are sufficiently similar, much empirical evidence suggests that OOD data can actually improve generalization error. We s how a counter-intuitive phenomenon: the generalization error of a task can be a non-monotonic function of the amount of OOD data. Specifically, we prove that ge neralization error can improve with small amounts of OOD data, and then get wors e than no OOD data with larger amounts. In other words, there is value in traini ng on small amounts of OOD data. We analytically demonstrate these results via F isher’s Linear Discriminant on synthetic datasets, and empirically demonstrate t hem via deep networks on computer vision benchmarks such as MNIST, CIFAR-10, CIN IC-10, PACS and DomainNet. In the idealistic setting where we know which samples are OOD, we show that these non-monotonic trends can be exploited using an appr opriately weighted objective of the target and OOD empirical risk. While its pra ctical utility is limited, this does suggest that if we can detect OOD samples, then there may be ways to benefit from them. When we do not know which samples a re OOD, we show how a number of go-to strategies such as data-augmentation, hype r-parameter optimization and pre-training are not enough to ensure that the targ et generalization error does not deteriorate with the number of OOD samples in t he dataset.

****

High Fidelity Image Counterfactuals with Probabilistic Causal Models

Fabio De Sousa Ribeiro, Tian Xia, Miguel Monteiro, Nick Pawlowski, Ben Glocker

We present a general causal generative modelling framework for accurate estimati on of high fidelity image counterfactuals with deep structural causal models. Es timation of interventional and counterfactual queries for high-dimensional struc tured variables, such as images, remains a challenging task. We leverage ideas f rom causal mediation analysis and advances in generative modelling to design new deep causal mechanisms for structured variables in causal models. Our experimen ts demonstrate that our proposed mechanisms are capable of accurate abduction an d estimation of direct, indirect and total effects as measured by axiomatic soun dness of counterfactuals.

****

Learning Noisy OR Bayesian Networks with Max-Product Belief Propagation

Antoine Dedieu, Guangyao Zhou, Dileep George, Miguel Lazaro-Gredilla

Noisy-OR Bayesian Networks (BNs) are a family of probabilistic graphical models which express rich statistical dependencies in binary data. Variational inferenc e (VI) has been the main method proposed to learn noisy-OR BNs with complex late nt structures (Jaakkola & Jordan, 1999; Ji et al., 2020; Buhai et al., 2020). Ho wever, the proposed VI approaches either (a) use a recognition network with stan dard amortized inference that cannot induce "explaining-away"; or (b) assume a s imple mean-field (MF) posterior which is vulnerable to bad local optima. Existin g MF VI methods also update the MF parameters sequentially which makes them inhe rently slow. In this paper, we propose parallel max-product as an alternative al gorithm for learning noisy-OR BNs with complex latent structures and we derive a

fast stochastic training scheme that scales to large datasets. We evaluate both approaches on several benchmarks where VI is the state-of-the-art and show that our method (a) achieves better test performance than Ji et al. (2020) for learning noisy-OR BNs with hierarchical latent structures on large sparse real datasets; (b) recovers a higher number of ground truth parameters than Buhai et al. (2020) from cluttered synthetic scenes; and (c) solves the 2D blind deconvolution problem from Lazaro-Gredilla et al. (2021) and variants - including binary matrix factorization - while VI catastrophically fails and is up to two orders of magnitude slower.

****

Learning-Rate-Free Learning by D-Adaptation

Aaron Defazio, Konstantin Mishchenko

The speed of gradient descent for convex Lipschitz functions is highly dependent on the choice of learning rate. Setting the learning rate to achieve the optimal convergence rate requires knowing the distance D from the initial point to the solution set. In this work, we describe a single-loop method, with no back-tracking or line searches, which does not require knowledge of D yet asymptotically achieves the optimal rate of convergence for the complexity class of convex Lipschitz functions. Our approach is the first parameter-free method for this class without additional multiplicative log factors in the convergence rate. We present extensive experiments for SGD and Adam variants of our method, where the method automatically matches hand-tuned learning rates across more than a dozen diverse machine learning problems, including large-scale vision and language problems. Our method is practical, efficient and requires no additional function value or gradient evaluations each step. An implementation is provided in the supplementary material.

****

Scaling Vision Transformers to 22 Billion Parameters

Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, Rodolphe Jenatton, Lucas Beyer, Michael Tschannen, Anurag Arnab, Xiao Wang, Carlos Riquelme Ruiz, Matthias Minderer, Joan Puigcerver, Utku Evci, Manoj Kumar, Sjoerd Van Steenkiste, Gamaleldin Fathy Elsayed, Aravindh Mahendran, Fisher Yu, Avital Oliver, Fantine Huot, Jasmijn Bastings, Mark Collier, Alexey A. Gritsenko, Vighnesh Birodkar, Cristina Nader Vasconcelos, Yi Tay, Thomas Mensink, Alexander Kolesnikov, Filip Pavetic, Dustin Tran, Thomas Kipf, Mario Lucic, Xiaohua Zhai, Daniel Keysers, Jeremiah J. Harmsen, Neil Houlsby

The scaling of Transformers has driven breakthrough capabilities for language models. At present, the largest large language models (LLMs) contain upwards of 100B parameters. Vision Transformers (ViT) have introduced the same architecture to image and video modelling, but these have not yet been successfully scaled to nearly the same degree; the largest dense ViT contains 4B parameters (Chen et al., 2022). We present a recipe for highly efficient and stable training of a 22B-parameter ViT (ViT-22B) and perform a wide variety of experiments on the resulting model. When evaluated on downstream tasks (often with a lightweight linear model on frozen features), ViT-22B demonstrates increasing performance with scale. We further observe other interesting benefits of scale, including an improved tradeoff between fairness and performance, state-of-the-art alignment to human visual perception in terms of shape/texture bias, and improved robustness. ViT-22B demonstrates the potential for "LLM-like" scaling in vision, and provides key steps towards getting there.

****

Efficient Bound of Lipschitz Constant for Convolutional Layers by Gram Iteration

Blaise Delattre, Quentin Barthélemy, Alexandre Araujo, Alexandre Allauzen

Since the control of the Lipschitz constant has a great impact on the training stability, generalization, and robustness of neural networks, the estimation of this value is nowadays a real scientific challenge. In this paper we introduce a precise, fast, and differentiable upper bound for the spectral norm of convolutional layers using circulant matrix theory and a new alternative to the Power iteration. Called the Gram iteration, our approach exhibits a superlinear convergen

ce. First, we show through a comprehensive set of experiments that our approach outperforms other state-of-the-art methods in terms of precision, computational cost, and scalability. Then, it proves highly effective for the Lipschitz regularization of convolutional neural networks, with competitive results against concurrent approaches.

****

Blossom: an Anytime Algorithm for Computing Optimal Decision Trees

Emir Demirovi■, Emmanuel Hebrard, Louis Jean

We propose a simple algorithm to learn optimal decision trees of bounded depth. This algorithm is essentially an anytime version of the state-of-the-art dynamic programming approach. It has virtually no overhead compared to heuristic methods and is comparable to the best exact methods to prove optimality on most data sets. Experiments show that whereas existing exact methods hardly scale to deep trees, this algorithm learns trees comparable to standard heuristics without computational overhead, and can significantly improve their accuracy when given more computation time, even for deep trees.

****

Optimizing NOTEARS Objectives via Topological Swaps

Chang Deng, Kevin Bello, Bryon Aragam, Pradeep Kumar Ravikumar

Recently, an intriguing class of non-convex optimization problems has emerged in the context of learning directed acyclic graphs (DAGs). These problems involve minimizing a given loss or score function, subject to a non-convex continuous constraint that penalizes the presence of cycles in a graph. In this work, we delve into the optimality challenges associated with this class of non-convex programs. To address these challenges, we propose a bi-level algorithm that leverages the non-convex constraint in a novel way. The outer level of the algorithm optimizes over topological orders by iteratively swapping pairs of nodes within the topological order of a DAG. A key innovation of our approach is the development of an effective method for generating a set of candidate swapping pairs for each iteration. At the inner level, given a topological order, we utilize off-the-shelf solvers that can handle linear constraints. The key advantage of our proposed algorithm is that it is guaranteed to find a local minimum or a KKT point under weaker conditions compared to previous work and finds solutions with lower scores. Extensive experiments demonstrate that our method outperforms state-of-the-art approaches in terms of achieving a better score. Additionally, our method can also be used as a post-processing algorithm to significantly improve the score of other algorithms. Code implementing the proposed method is available at https://github.com/duntrain/topo.

****

Uncertainty Estimation by Fisher Information-based Evidential Deep Learning

Danruo Deng, Guangyong Chen, Yang Yu, Furui Liu, Pheng-Ann Heng

Uncertainty estimation is a key factor that makes deep learning reliable in practical applications. Recently proposed evidential neural networks explicitly account for different uncertainties by treating the network’s outputs as evidence to parameterize the Dirichlet distribution, and achieve impressive performance in uncertainty estimation. However, for high data uncertainty samples but annotated with the one-hot label, the evidence-learning process for those mislabeled classes is over-penalized and remains hindered. To address this problem, we propose a novel method, Fisher Information-based Evidential Deep Learning ($\mathcal{I}$-EDL). In particular, we introduce Fisher Information Matrix (FIM) to measure the informativeness of evidence carried by each sample, according to which we can dynamically reweight the objective loss terms to make the network more focus on the representation learning of uncertain classes. The generalization ability of our network is further improved by optimizing the PAC-Bayesian bound. As demonstrated empirically, our proposed method consistently outperforms traditional EDL-related algorithms in multiple uncertainty estimation tasks, especially in the more challenging few-shot classification settings.

****

Multi-channel Autobidding with Budget and ROI Constraints

Yuan Deng, Negin Golrezaei, Patrick Jaillet, Jason Cheuk Nam Liang, Vahab Mirrok

ni
In digital online advertising, advertisers procure ad impressions simultaneously on multiple platforms, or so-called channels, such as Google Ads, Meta Ads Manager, etc., each of which consists of numerous ad auctions. We study how an advertiser maximizes total conversion (e.g. ad clicks) while satisfying aggregate return-on-investment (ROI) and budget constraints across all channels. In practice, an advertiser does not have control over, and thus cannot globally optimize, which individual ad auctions she participates in for each channel, and instead authorizes a channel to procure impressions on her behalf: the advertiser can only utilize two levers on each channel, namely setting a per-channel budget and per-channel target ROI. In this work, we first analyze the effectiveness of each of these levers for solving the advertiser's global multi-channel problem. We show that when an advertiser only optimizes over per-channel ROIs, her total conversion can be arbitrarily worse than what she could have obtained in the global problem. Further, we show that the advertiser can achieve the global optimal conversion when she only optimizes over per-channel budgets. In light of this finding, under a bandit feedback setting that mimics real-world scenarios where advertisers have limited information on ad auctions in each channels and how channels procure ads, we present an efficient learning algorithm that produces per-channel budgets whose resulting conversion approximates that of the global optimal problem.
****
Surrogate Module Learning: Reduce the Gradient Error Accumulation in Training Spiking Neural Networks
Shikuang Deng, Hao Lin, Yuhang Li, Shi Gu
Spiking neural networks provide an alternative solution to conventional artificial neural networks with energy-saving and high-efficiency characteristics after hardware implantation. However, due to its non-differentiable activation function and the temporally delayed accumulation in outputs, the direct training of SNNs is extraordinarily tough even adopting a surrogate gradient to mimic the backpropagation. For SNN training, this non-differentiability causes the intrinsic gradient error that would be magnified through layerwise backpropagation, especially through multiple layers. In this paper, we propose a novel approach to reducing gradient error from a new perspective called surrogate module learning (SML). Surrogate module learning tries to construct a shortcut path to back-propagate more accurate gradient to a certain SNN part utilizing the surrogate modules. Then, we develop a new loss function for concurrently training the network and enhancing the surrogate modules' surrogate capacity. We demonstrate that when the outputs of surrogate modules are close to the SNN output, the fraction of the gradient error drops significantly. Our method consistently and significantly enhances the performance of SNNs on all experiment datasets, including CIFAR-10/100, ImageNet, and ES-ImageNet. For example, for spiking ResNet-34 architecture on ImageNet, we increased the SNN accuracy by 3.46%.
****
Confidence and Dispersity Speak: Characterizing Prediction Matrix for Unsupervised Accuracy Estimation
Weijian Deng, Yumin Suh, Stephen Gould, Liang Zheng
This work aims to assess how well a model performs under distribution shifts without using labels. While recent methods study prediction confidence, this work reports prediction dispersity is another informative cue. Confidence reflects whether the individual prediction is certain; dispersity indicates how the overall predictions are distributed across all categories. Our key insight is that a well-performing model should give predictions with high confidence and high dispersity. That is, we need to consider both properties so as to make more accurate estimates. To this end, we use nuclear norm that has been shown to be effective in characterizing both properties. Extensive experiments validate the effectiveness of nuclear norm for various models (e.g., ViT and ConvNeXt), different datasets (e.g., ImageNet and CUB-200), and diverse types of distribution shifts (e.g., style shift and reproduction shift). We show that nuclear norm is more accurate and robust in accuracy estimation than existing methods. Furthermore, we validat

e the feasibility of other measurements (e.g., mutual information maximization) for characterizing dispersity and confidence. Lastly, we investigate the limitation of the nuclear norm, study its improved variant under severe class imbalance, and discuss potential directions.
****
Great Models Think Alike: Improving Model Reliability via Inter-Model Latent Agreement
Ailin Deng, Miao Xiong, Bryan Hooi
Reliable application of machine learning is of primary importance to the practical deployment of deep learning methods. A fundamental challenge is that models are often unreliable due to overconfidence. In this paper, we estimate a model's reliability by measuring the agreement between its latent space, and the latent space of a foundation model. However, it is challenging to measure the agreement between two different latent spaces due to their incoherence, e.g., arbitrary rotations and different dimensionality. To overcome this incoherence issue, we design a neighborhood agreement measure between latent spaces and find that this agreement is surprisingly well-correlated with the reliability of a model's predictions. Further, we show that fusing neighborhood agreement into a model's predictive confidence in a post-hoc way significantly improves its reliability. Theoretical analysis and extensive experiments on failure detection across various datasets verify the effectiveness of our method on both in-distribution and out-of-distribution settings.
****
Hyperbolic Image-text Representations
Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, Shanmukha Ramakrishna Vedantam
Visual and linguistic concepts naturally organize themselves in a hierarchy, where a textual concept "dog" entails all images that contain dogs. Despite being intuitive, current large-scale vision and language models such as CLIP do not explicitly capture such hierarchy. We propose MERU, a contrastive model that yields hyperbolic representations of images and text. Hyperbolic spaces have suitable geometric properties to embed tree-like data, so MERU can better capture the underlying hierarchy in image-text datasets. Our results show that MERU learns a highly interpretable and structured representation space while being competitive with CLIP's performance on standard multi-modal tasks like image classification and image-text retrieval.
****
Hardware-Aware Compression with Random Operation Access Specific Tile (ROAST) Hashing
Aditya Desai, Keren Zhou, Anshumali Shrivastava
Advancements in deep learning are often associated with increasing model sizes. Training and deploying large models require sophisticated hardware and incur significantly higher costs. Thus, model compression is a widely explored approach to solving the problem. However, SOTA techniques fall short in one or more desirable aspects of compression - for instance, pruning does not reduce memory for training, quantization can only provide up to 32$\times$ compression, HashedNet is cache-inefficient, etc. This paper proposes a model-agnostic, cache-friendly, and hardware-aware model compression approach: Random Operation Access Specific Tile (ROAST) hashing. ROAST collapses the parameters by clubbing them through a lightweight mapping. While clubbing these parameters, ROAST utilizes cache hierarchies by aligning the memory access pattern with the parameter access pattern. ROAST is up to ${\sim}25\times$ faster to train and ${\sim}50\times$ faster to infer than the popular parameter sharing method HashedNet. Additionally, ROAST introduces global weight sharing, which is empirically and theoretically superior to local weight sharing in HashedNet, and can be of independent interest. With ROAST, we can efficiently train and deploy the model using a much smaller memory footprint ($\sim 10 - 100\times$ lesser) in text and image classification tasks. ROAST-MM kernel implementation is open-source (https://github.com/apd10/RzLinear/tree/stable)
****

The case for 4-bit precision: k-bit Inference Scaling Laws
Tim Dettmers, Luke Zettlemoyer
Quantization methods reduce the number of bits required to represent each parame ter in a model, trading accuracy for smaller memory footprints and inference lat encies. However, the final model size depends on both the number of parameters o f the original model and the rate of compression. For example, a 30B 8-bit model and a 60B 4-bit model have the same number of bits but may have very different zero-shot accuracies. In this work, we study this trade-off by developing infere nce scaling laws of zero-shot performance in Large Language Models (LLMs) to det ermine the bit-precision and model size that maximizes zero-shot performance. We run more than 35,000 experiments with 16-bit inputs and k-bit parameters to exa mine which zero-shot quantization methods improve scaling for 3 to 8-bit precisi on at scales of 19M to 176B parameters across the LLM families BLOOM, OPT, NeoX/ Pythia, and GPT-2. We find that it is challenging to improve the bit-level scali ng trade-off, with the only improvements being the use of a small block size – s plitting the parameters into small independently quantized blocks – and the quan tization data type being used (e.g., Int vs Float). Overall, our findings show t hat 4-bit precision is almost universally optimal for total model bits and zero- shot accuracy.
****
Fairness in Matching under Uncertainty
Siddartha Devic, David Kempe, Vatsal Sharan, Aleksandra Korolova
The prevalence and importance of algorithmic two-sided marketplaces has drawn at tention to the issue of fairness in such settings. Algorithmic decisions are use d in assigning students to schools, users to advertisers, and applicants to job interviews. These decisions should heed the preferences of individuals, and simu ltaneously be fair with respect to their merits (synonymous with fit, future per formance, or need). Merits conditioned on observable features are always uncerta in, a fact that is exacerbated by the widespread use of machine learning algorit hms to infer merit from the observables. As our key contribution, we carefully a xiomatize a notion of individual fairness in the two-sided marketplace setting w hich respects the uncertainty in the merits; indeed, it simultaneously recognize s uncertainty as the primary potential cause of unfairness and an approach to ad dress it. We design a linear programming framework to find fair utility-maximizi ng distributions over allocations, and we show that the linear program is robust to perturbations in the estimated parameters of the uncertain merit distributio ns, a key property in combining the approach with machine learning techniques.
****
Efficient Parametric Approximations of Neural Network Function Space Distance
Nikita Dhawan, Sicong Huang, Juhan Bae, Roger Baker Grosse
It is often useful to compactly summarize important properties of model paramete rs and training data so that they can be used later without storing and/or itera ting over the entire dataset. As a specific case, we consider estimating the Fun ction Space Distance (FSD) over a training set, i.e. the average discrepancy bet ween the outputs of two neural networks. We propose a Linearized Activation Func tion TRick (LAFTR) and derive an efficient approximation to FSD for ReLU neural networks. The key idea is to approximate the architecture as a linear network wi th stochastic gating. Despite requiring only one parameter per unit of the netwo rk, our approach outcompetes other parametric approximations with larger memory requirements. Applied to continual learning, our parametric approximation is com petitive with state-of-the-art nonparametric approximations, which require stori ng many training examples. Furthermore, we show its efficacy in estimating influ ence functions accurately and detecting mislabeled examples without expensive it erations over the entire dataset.
****
A Large-Scale Study of Probabilistic Calibration in Neural Network Regression
Victor Dheur, Souhaib Ben Taieb
Accurate probabilistic predictions are essential for optimal decision making. Wh ile neural network miscalibration has been studied primarily in classification, we investigate this in the less-explored domain of regression. We conduct the la

rgest empirical study to date to assess the probabilistic calibration of neural networks. We also analyze the performance of recalibration, conformal, and regularization methods to enhance probabilistic calibration. Additionally, we introduce novel differentiable recalibration and regularization methods, uncovering new insights into their effectiveness. Our findings reveal that regularization methods offer a favorable tradeoff between calibration and sharpness. Post-hoc methods exhibit superior probabilistic calibration, which we attribute to the finite-sample coverage guarantee of conformal prediction. Furthermore, we demonstrate that quantile recalibration can be considered as a specific case of conformal prediction. Our study is fully reproducible and implemented in a common code base for fair comparisons.

****

Nearly Minimax Optimal Regret for Learning Linear Mixture Stochastic Shortest Path

Qiwei Di, Jiafan He, Dongruo Zhou, Quanquan Gu

We study the Stochastic Shortest Path (SSP) problem with a linear mixture transition kernel, where an agent repeatedly interacts with a stochastic environment and seeks to reach certain goal state while minimizing the cumulative cost. Existing works often assume a strictly positive lower bound of the cost function or an upper bound of the expected length for the optimal policy. In this paper, we propose a new algorithm to eliminate these restrictive assumptions. Our algorithm is based on extended value iteration with a fine-grained variance-aware confidence set, where the variance is estimated recursively from high-order moments. Our algorithm achieves an $\tilde{\mathcal{O}}(dB_*\sqrt{K})$ regret bound, where d is the dimension of the feature mapping in the linear transition kernel, B_* is the upper bound of the total cumulative cost for the optimal policy, and K is the number of episodes. Our regret upper bound matches the $\Omega(dB_*\sqrt{K})$ lower bound of linear mixture SSPs in Min et al. (2022), which suggests that our algorithm is nearly minimax optimal.

****

On Over-Squashing in Message Passing Neural Networks: The Impact of Width, Depth, and Topology

Francesco Di Giovanni, Lorenzo Giusti, Federico Barbero, Giulia Luise, Pietro Lio, Michael M. Bronstein

Message Passing Neural Networks (MPNNs) are instances of Graph Neural Networks that leverage the graph to send messages over the edges. This inductive bias leads to a phenomenon known as over-squashing, where a node feature is insensitive to information contained at distant nodes. Despite recent methods introduced to mitigate this issue, an understanding of the causes for over-squashing and of possible solutions are lacking. In this theoretical work, we prove that: (i) Neural network width can mitigate over-squashing, but at the cost of making the whole network more sensitive; (ii) Conversely, depth cannot help mitigate over-squashing: increasing the number of layers leads to over-squashing being dominated by vanishing gradients; (iii) The graph topology plays the greatest role, since over-squashing occurs between nodes at high commute time. Our analysis provides a unified framework to study different recent methods introduced to cope with over-squashing and serves as a justification for a class of methods that fall under graph rewiring.

****

Nearly-Linear Time and Streaming Algorithms for Outlier-Robust PCA

Ilias Diakonikolas, Daniel Kane, Ankit Pensia, Thanasis Pittas

We study principal component analysis (PCA), where given a dataset in $\mathbb R^d$ from a distribution, the task is to find a unit vector v that approximately maximizes the variance of the distribution after being projected along v. Despite being a classical task, standard estimators fail drastically if the data contains even a small fraction of outliers, motivating the problem of robust PCA. Recent work has developed computationally-efficient algorithms for robust PCA that either take super-linear time or have sub-optimal error guarantees. Our main contribution is to develop a nearly linear time algorithm for robust PCA with near-optimal error guarantees. We also develop a single-pass streaming algorithm

for robust PCA with memory usage nearly-linear in the dimension.
****
Near-Optimal Cryptographic Hardness of Agnostically Learning Halfspaces and ReLU Regression under Gaussian Marginals
Ilias Diakonikolas, Daniel Kane, Lisheng Ren
We study the task of agnostically learning halfspaces under the Gaussian distribution. Specifically, given labeled examples $(\\mathbf{x},y)$ from an unknown distribution on $\\mathbb{R}^n \\times \\{\pm 1 \\}$, whose marginal distribution on $\\mathbf{x}$ is the standard Gaussian and the labels y can be arbitrary, the goal is to output a hypothesis with 0-1 loss $\\mathrm{OPT}+\\epsilon$, where $\\mathrm{OPT}$ is the 0-1 loss of the best-fitting halfspace. We prove a near-optimal computational hardness result for this task, under the widely believed sub-exponential time hardness of the Learning with Errors (LWE) problem. Prior hardness results are either qualitatively suboptimal or apply to restricted families of algorithms. Our techniques extend to yield near-optimal lower bounds for related problems, including ReLU regression.
****
Improving Graph Generation by Restricting Graph Bandwidth
Nathaniel Lee Diamant, Alex M Tseng, Kangway V. Chuang, Tommaso Biancalani, Gabriele Scalia
Deep graph generative modeling has proven capable of learning the distribution of complex, multi-scale structures characterizing real-world graphs. However, one of the main limitations of existing methods is their large output space, which limits generation scalability and hinders accurate modeling of the underlying distribution. To overcome these limitations, we propose a novel approach that significantly reduces the output space of existing graph generative models. Specifically, starting from the observation that many real-world graphs have low graph bandwidth, we restrict graph bandwidth during training and generation. Our strategy improves both generation scalability and quality without increasing architectural complexity or reducing expressiveness. Our approach is compatible with existing graph generative methods, and we describe its application to both autoregressive and one-shot models. We extensively validate our strategy on synthetic and real datasets, including molecular graphs. Our experiments show that, in addition to improving generation efficiency, our approach consistently improves generation quality and reconstruction accuracy. The implementation is made available.
****
Forward-Backward Gaussian Variational Inference via JKO in the Bures-Wasserstein Space
Michael Ziyang Diao, Krishna Balasubramanian, Sinho Chewi, Adil Salim
Variational inference (VI) seeks to approximate a target distribution π by an element of a tractable family of distributions. Of key interest in statistics and machine learning is Gaussian VI, which approximates π by minimizing the Kullback-Leibler (KL) divergence to π over the space of Gaussians. In this work, we develop the (Stochastic) Forward-Backward Gaussian Variational Inference (FB-GVI) algorithm to solve Gaussian VI. Our approach exploits the composite structure of the KL divergence, which can be written as the sum of a smooth term (the potential) and a non-smooth term (the entropy) over the Bures-Wasserstein (BW) space of Gaussians endowed with the Wasserstein distance. For our proposed algorithm, we obtain state-of-the-art convergence guarantees when π is log-smooth and log-concave, as well as the first convergence guarantees to first-order stationary solutions when π is only log-smooth.
****
Subset-Based Instance Optimality in Private Estimation
Travis Dick, Alex Kulesza, Ziteng Sun, Ananda Theertha Suresh
We propose a new definition of instance optimality for differentially private estimation algorithms. Our definition requires an optimal algorithm to compete, simultaneously for every dataset D, with the best private benchmark algorithm that (a) knows D in advance and (b) is evaluated by its worst-case performance on large subsets of D. That is, the benchmark algorithm need not perform well when potentially extreme points are added to D; it only has to handle the remov

al of a small number of real data points that already exist. This makes our benc
hmark significantly stronger than those proposed in prior work. We nevertheless
show, for real-valued datasets, how to construct private algorithms that achieve
 our notion of instance optimality when estimating a broad class of dataset prop
erties, including means, quantiles, and ℓ_p-norm minimizers. For means in p
articular, we provide a detailed analysis and show that our algorithm simultaneo
usly matches or exceeds the asymptotic performance of existing algorithms under
a range of distributional assumptions.
****
Pareto Manifold Learning: Tackling multiple tasks via ensembles of single-task m
odels
Nikolaos Dimitriadis, Pascal Frossard, François Fleuret
In Multi-Task Learning (MTL), tasks may compete and limit the performance achiev
ed on each other, rather than guiding the optimization to a solution, superior t
o all its single-task trained counterparts. Since there is often not a unique so
lution optimal for all tasks, practitioners have to balance tradeoffs between ta
sks' performance, and resort to optimality in the Pareto sense. Most MTL methodo
logies either completely neglect this aspect, and instead of aiming at learning
a Pareto Front, produce one solution predefined by their optimization schemes, o
r produce diverse but discrete solutions. Recent approaches parameterize the Par
eto Front via neural networks, leading to complex mappings from tradeoff to obje
ctive space. In this paper, we conjecture that the Pareto Front admits a linear
parameterization in parameter space, which leads us to propose Pareto Manifold L
earning, an ensembling method in weight space. Our approach produces a continuou
s Pareto Front in a single training run, that allows to modulate the performance
 on each task during inference. Experiments on multi-task learning benchmarks, r
anging from image classification to tabular datasets and scene understanding, sh
ow that Pareto Manifold Learning outperforms state-of-the-art single-point algor
ithms, while learning a better Pareto parameterization than multi-point baseline
s.
****
Bayesian Reparameterization of Reward-Conditioned Reinforcement Learning with En
ergy-based Models
Wenhao Ding, Tong Che, Ding Zhao, Marco Pavone
Recently, reward-conditioned reinforcement learning (RCRL) has gained popularity
 due to its simplicity, flexibility, and off-policy nature. However, we will sho
w that current RCRL approaches are fundamentally limited and fail to address two
 critical challenges of RCRL – improving generalization on high reward-to-go (RT
G) inputs, and avoiding out-of-distribution (OOD) RTG queries during testing tim
e. To address these challenges when training vanilla RCRL architectures, we prop
ose Bayesian Reparameterized RCRL (BR-RCRL), a novel set of inductive biases for
 RCRL inspired by Bayes' theorem. BR-RCRL removes a core obstacle preventing van
illa RCRL from generalizing on high RTG inputs – a tendency that the model treat
s different RTG inputs as independent values, which we term "RTG Independence".
BR-RCRL also allows us to design an accompanying adaptive inference method, whic
h maximizes total returns while avoiding OOD queries that yield unpredictable be
haviors in vanilla RCRL methods. We show that BR-RCRL achieves state-of-the-art
performance on the Gym-Mujoco and Atari offline RL benchmarks, improving upon va
nilla RCRL by up to 11%.
****
DSGD-CECA: Decentralized SGD with Communication-Optimal Exact Consensus Algorith
m
Lisang Ding, Kexin Jin, Bicheng Ying, Kun Yuan, Wotao Yin
Decentralized Stochastic Gradient Descent (SGD) is an emerging neural network tr
aining approach that enables multiple agents to train a model collaboratively an
d simultaneously. Rather than using a central parameter server to collect gradie
nts from all the agents, each agent keeps a copy of the model parameters and com
municates with a small number of other agents to exchange model updates. Their c
ommunication, governed by the communication topology and gossip weight matrices,
 facilitates the exchange of model updates. The state-of-the-art approach uses t

he dynamic one-peer exponential-2 topology, achieving faster training times and improved scalability than the ring, grid, torus, and hypercube topologies. However, this approach requires a power-of-2 number of agents, which is impractical at scale. In this paper, we remove this restriction and propose Decentralized SGD with Communication-optimal Exact Consensus Algorithm (DSGD-CECA), which works for any number of agents while still achieving state-of-the-art properties. In particular, DSGD-CECA incurs a unit per-iteration communication overhead and an $\tilde{O}(n^3)$ transient iteration complexity. Our proof is based on newly discovered properties of gossip weight matrices and a novel approach to combine them with DSGD's convergence analysis. Numerical experiments show the efficiency of DSGD-CECA.
****

Open-Vocabulary Universal Image Segmentation with MaskCLIP
Zheng Ding, Jieke Wang, Zhuowen Tu
In this paper, we tackle an emerging computer vision task, open-vocabulary universal image segmentation, that aims to perform semantic/instance/panoptic segmentation (background semantic labeling + foreground instance segmentation) for arbitrary categories of text-based descriptions in inference time. We first build a baseline method by directly adopting pre-trained CLIP models without finetuning or distillation. We then develop MaskCLIP, a Transformer-based approach with a MaskCLIP Visual Encoder, which is an encoder-only module that seamlessly integrates mask tokens with a pre-trained ViT CLIP model for semantic/instance segmentation and class prediction. MaskCLIP learns to efficiently and effectively utilize pre-trained partial/dense CLIP features within the MaskCLIP Visual Encoder that avoids the time-consuming student-teacher training process. MaskCLIP outperforms previous methods for semantic/instance/panoptic segmentation on ADE20K and PASCAL datasets. We show qualitative illustrations for MaskCLIP with online custom categories. Project website: https://maskclip.github.io.
****

Entity Divider with Language Grounding in Multi-Agent Reinforcement Learning
Ziluo Ding, Wanpeng Zhang, Junpeng Yue, Xiangjun Wang, Tiejun Huang, Zongqing Lu
We investigate the use of natural language to drive the generalization of policies in multi-agent settings. Unlike single-agent settings, the generalization of policies should also consider the influence of other agents. Besides, with the increasing number of entities in multi-agent settings, more agent-entity interactions are needed for language grounding, and the enormous search space could impede the learning process. Moreover, given a simple general instruction, e.g., beating all enemies, agents are required to decompose it into multiple subgoals and figure out the right one to focus on. Inspired by previous work, we try to address these issues at the entity level and propose a novel framework for language grounding in multi-agent reinforcement learning, entity divider (EnDi). EnDi enables agents to independently learn subgoal division at the entity level and act in the environment based on the associated entities. The subgoal division is regularized by agent modeling to avoid subgoal conflicts and promote coordinated strategies. Empirically, EnDi demonstrates the strong generalization ability to unseen games with new dynamics and expresses the superiority over existing methods. The code is available at https://github.com/PKU-RL/EnDi.
****

PixelAsParam: A Gradient View on Diffusion Sampling with Guidance
Anh-Dung Dinh, Daochang Liu, Chang Xu
Diffusion models recently achieved state-of-the-art in image generation. They mainly utilize the denoising framework, which leverages the Langevin dynamics process for image sampling. Recently, the guidance method has modified this process to add conditional information to achieve a controllable generator. However, the current guidance on denoising processes suffers from the trade-off between diversity, image quality, and conditional information. In this work, we propose to view this guidance sampling process from a gradient view, where image pixels are treated as parameters being optimized, and each mathematical term in the sampling process represents one update direction. This perspective reveals more insights into the conflict problems between updated directions on the pixels, which cau

se the trade-off as mentioned previously. We investigate the conflict problems a nd propose to solve them by a simple projection method. The experimental results evidently improve over different baselines on datasets with various resolutions .

****

Second-Order Optimization with Lazy Hessians

Nikita Doikov, El Mahdi Chayti, Martin Jaggi

We analyze Newton’s method with lazy Hessian updates for solving general possibl y non-convex optimization problems. We propose to reuse a previously seen Hessia n for several iterations while computing new gradients at each step of the metho d. This significantly reduces the overall arithmetic complexity of second-order optimization schemes. By using the cubic regularization technique, we establish fast global convergence of our method to a second-order stationary point, while the Hessian does not need to be updated each iteration. For convex problems, we justify global and local superlinear rates for lazy Newton steps with quadratic regularization, which is easier to compute. The optimal frequency for updating t he Hessian is once every d iterations, where d is the dimension of the probl em. This provably improves the total arithmetic complexity of second-order algor ithms by a factor $\sqrt{d}$.

****

Polynomial Preconditioning for Gradient Methods

Nikita Doikov, Anton Rodomanov

We study first-order methods with preconditioning for solving structured convex optimization problems. We propose a new family of preconditioners generated by t he symmetric polynomials. They provide the first-order optimization methods with a provable improvement of the condition number, cutting the gaps between highes t eigenvalues, without explicit knowledge of the actual spectrum. We give a stoc hastic interpretation of this preconditioning in terms of the coordinate volume sampling and compare it with other classical approaches, including the Chebyshev polynomials. We show how to incorporate a polynomial preconditioning into the G radient and Fast Gradient Methods and establish their better global complexity b ounds. Finally, we propose a simple adaptive search procedure that automatically ensures the best polynomial preconditioning for the Gradient Method, minimizing the objective along a low-dimensional Krylov subspace. Numerical experiments co nfirm the efficiency of our preconditioning strategies for solving various machi ne learning problems.

****

On Data Manifolds Entailed by Structural Causal Models

Ricardo Dominguez-Olmedo, Amir-Hossein Karimi, Georgios Arvanitidis, Bernhard Sc hölkopf

The geometric structure of data is an important inductive bias in machine learni ng. In this work, we characterize the data manifolds entailed by structural caus al models. The strengths of the proposed framework are twofold: firstly, the geo metric structure of the data manifolds is causally informed, and secondly, it en ables causal reasoning about the data manifolds in an interventional and a count erfactual sense. We showcase the versatility of the proposed framework by applyi ng it to the generation of causally-grounded counterfactual explanations for mac hine learning classifiers, measuring distances along the data manifold in a diff erential geometric-principled manner.

****

Towards Understanding and Reducing Graph Structural Noise for GNNs

Mingze Dong, Yuval Kluger

Graph neural networks (GNNs) have emerged as a powerful paradigm to learn from r elational data mostly through applying the message passing mechanism. However, t his approach may exhibit suboptimal performance when applied to graphs possessin g various structural issues. In this work, we focus on understanding and allevia ting the effect of graph structural noise on GNN performance. To evaluate the gr aph structural noise in real data, we propose edge signal-to-noise ratio (ESNR), a novel metric evaluating overall edge noise level with respect to data feature s or labels based on random matrix theory. We have found striking concordance be

tween the proposed ESNR metric and the GNN performance in various simulated and real data. To reduce the effect of the noise, we propose GPS (Graph Propensity Score) graph rewiring, which estimates the edge likelihood for rewiring data graphs based on self-supervised link prediction. We provide a theoretical guarantee for GPS graph rewiring and demonstrate its efficacy by comprehensive benchmarks.
****
SpeedDETR: Speed-aware Transformers for End-to-end Object Detection
Peiyan Dong, Zhenglun Kong, Xin Meng, Peng Zhang, Hao Tang, Yanzhi Wang, Chih-Hsien Chou
Vision Transformers (ViTs) have continuously achieved new milestones in object detection. However, the considerable computation and memory burden compromise their efficiency and generalization of deployment on resource-constraint devices. Besides, efficient transformer-based detectors designed by existing works can hardly achieve a realistic speedup, especially on multi-core processors (e.g., GPUs). The main issue is that the current literature solely concentrates on building algorithms with minimal computation, oblivious that the practical latency can also be affected by the memory access cost and the degree of parallelism. Therefore, we propose SpeedDETR, a novel speed-aware transformer for end-to-end object detectors, achieving high-speed inference on multiple devices. Specifically, we design a latency prediction model which can directly and accurately estimate the network latency by analyzing network properties, hardware memory access pattern, and degree of parallelism. Following the effective local-to-global visual modeling process and the guidance of the latency prediction model, we build our hardware-oriented architecture design and develop a new family of SpeedDETR. Experiments on the MS COCO dataset show SpeedDETR outperforms current DETR-based methods on Tesla V100. Even acceptable speed inference can be achieved on edge GPUs.
****
Understand and Modularize Generator Optimization in ELECTRA-style Pretraining
Chengyu Dong, Liyuan Liu, Hao Cheng, Jingbo Shang, Jianfeng Gao, Xiaodong Liu
Despite the effectiveness of ELECTRA-style pre-training, their performance is dependent on the careful selection of the model size for the auxiliary generator, leading to high trial-and-error costs. In this paper, we present the first systematic study of this problem. Our theoretical investigation highlights the importance of controlling the generator capacity in ELECTRA-style training. Meanwhile, we found it is not handled properly in the original ELECTRA design, leading to the sensitivity issue. Specifically, since adaptive optimizers like Adam will cripple the weighing of individual losses in the joint optimization, the original design fails to control the generator training effectively. To regain control over the generator, we modularize the generator optimization by decoupling the generator optimizer and discriminator optimizer completely, instead of simply relying on the weighted objective combination. Our simple technique reduced the sensitivity of ELECTRA training significantly and obtains considerable performance gain compared to the original design.
****
Diversity-enhancing Generative Network for Few-shot Hypothesis Adaptation
Ruijiang Dong, Feng Liu, Haoang Chi, Tongliang Liu, Mingming Gong, Gang Niu, Masashi Sugiyama, Bo Han
Generating unlabeled data has been recently shown to help address the few-shot hypothesis adaptation (FHA) problem, where we aim to train a classifier for the target domain with a few labeled target-domain data and a well-trained source-domain classifier (i.e., a source hypothesis), for the additional information of the highly-compatible unlabeled data. However, the generated data of the existing methods are extremely similar or even the same. The strong dependency among the generated data will lead the learning to fail. In this paper, we propose a diversity-enhancing generative network (DEG-Net) for the FHA problem, which can generate diverse unlabeled data with the help of a kernel independence measure: the Hilbert-Schmidt independence criterion (HSIC). Specifically, DEG-Net will generate data via minimizing the HSIC value (i.e., maximizing the independence) among the semantic features of the generated data. By DEG-Net, the generated unlabeled data are more diverse and more effective for addressing the FHA problem. Experim

ental results show that the DEG-Net outperforms existing FHA baselines and furth
er verifies that generating diverse data plays an important role in addressing t
he FHA problem.
****

PASTA: Pessimistic Assortment Optimization
Juncheng Dong, Weibin Mo, Zhengling Qi, Cong Shi, Ethan X Fang, Vahid Tarokh
We consider a fundamental class of assortment optimization problems in an offlin
e data-driven setting. The firm does not know the underlying customer choice mod
el but has access to an offline dataset consisting of the historically offered a
ssortment set, customer choice, and revenue. The objective is to use the offline
 dataset to find an optimal assortment. Due to the combinatorial nature of assor
tment optimization, the problem of insufficient data coverage is likely to occur
 in the offline dataset. Therefore, designing a provably efficient offline learn
ing algorithm becomes a significant challenge. To this end, based on the princip
le of pessimism, we propose a novel algorithm called Pessimistic ASsortment opTi
mizAtion (PASTA for short), which can correctly identify the optimal assortment
by only requiring the offline data to cover the optimal assortment under general
 settings. In particular, we establish the first regret bound for the offline as
sortment optimization problem under the celebrated multinomial logit model (MNL)
. We also propose an efficient computational procedure to solve our pessimistic
assortment optimization problem. Our numerical studies demonstrate the superiori
ty of the proposed method over the existing baseline method.
****

Adaptively Weighted Data Augmentation Consistency Regularization for Robust Opti
mization under Concept Shift
Yijun Dong, Yuege Xie, Rachel Ward
Concept shift is a prevailing problem in natural tasks like medical image segmen
tation where samples usually come from different subpopulations with variant cor
relations between features and labels. One common type of concept shift in medic
al image segmentation is the "information imbalance" between label-sparse sample
s with few (if any) segmentation labels and label-dense samples with plentiful l
abeled pixels. Existing distributionally robust algorithms have focused on adapt
ively truncating/down-weighting the "less informative" (i.e., label-sparse in ou
r context) samples. To exploit data features of label-sparse samples more effici
ently, we propose an adaptively weighted online optimization algorithm — AdaWAC
— to incorporate data augmentation consistency regularization in sample reweight
ing. Our method introduces a set of trainable weights to balance the supervised
loss and unsupervised consistency regularization of each sample separately. At t
he saddle point of the underlying objective, the weights assign label-dense samp
les to the supervised loss and label-sparse samples to the unsupervised consiste
ncy regularization. We provide a convergence guarantee by recasting the optimiza
tion as online mirror descent on a saddle point problem. Our empirical results d
emonstrate that AdaWAC not only enhances the segmentation performance and sample
 efficiency but also improves the robustness to concept shift on various medical
 image segmentation tasks with different UNet-style backbones.
****

Does Sparsity Help in Learning Misspecified Linear Bandits?
Jialin Dong, Lin Yang
Recently, the study of linear misspecified bandits has generated intriguing impl
ications of the hardness of learning in bandits and reinforcement learning (RL).
 In particular, Du et al. (2020) shows that even if a learner is given linear fe
atures in $\mathbb{R}^d$ that approximate the rewards in a bandit or RL with a u
niform error of ε, searching for an $O(\varepsilon)$-optimal action
requires pulling at least $\Omega(\exp(d))$ queries. Furthermore, Lattimore et a
l. (2020) show that a degraded $O(\varepsilon\sqrt{d})$-optimal solution can be
learned within $\operatorname{poly}(d/\varepsilon)$ queries. Yet it is unknown w
hether a structural assumption on the ground-truth parameter, such as sparsity,
could break $\varepsilon\sqrt{d}$ barrier. In this paper, we address this questi
on by showing that algorithms can obtain $O(\varepsilon)$-optimal actions by que
rying $\tilde{O}(\exp(m\varepsilon))$ actions, where m is the sparsity paramet

er, removing the $\exp(d)$-dependence. We further show (with an information-theo
retical lower bound) that this is the best possible if one demands an error $ m^
{\delta}\varepsilon$ for $0<\delta<1$. We further show that $\operatorname{poly}
(m/\varepsilon)$ bounds are possible when the linear features are "good”. These
results provide a nearly complete picture of how sparsity can help in misspecifi
ed bandit learning and provide a deeper understanding of when linear features ar
e “useful” for bandit and reinforcement learning with misspecification.
****

Symmetry-Aware Robot Design with Structured Subgroups
Heng Dong, Junyu Zhang, Tonghan Wang, Chongjie Zhang
Robot design aims at learning to create robots that can be easily controlled and
 perform tasks efficiently. Previous works on robot design have proven its abili
ty to generate robots for various tasks. However, these works searched the robot
s directly from the vast design space and ignored common structures, resulting i
n abnormal robots and poor performance. To tackle this problem, we propose a Sym
metry-Aware Robot Design (SARD) framework that exploits the structure of the des
ign space by incorporating symmetry searching into the robot design process. Spe
cifically, we represent symmetries with the subgroups of the dihedral group and
search for the optimal symmetry in structured subgroups. Then robots are designe
d under the searched symmetry. In this way, SARD can design efficient symmetric
robots while covering the original design space, which is theoretically analyzed
. We further empirically evaluate SARD on various tasks, and the results show it
s superior efficiency and generalizability.
****

DoCoFL: Downlink Compression for Cross-Device Federated Learning
Ron Dorfman, Shay Vargaftik, Yaniv Ben-Itzhak, Kfir Yehuda Levy
Many compression techniques have been proposed to reduce the communication overh
ead of Federated Learning training procedures. However, these are typically desi
gned for compressing model updates, which are expected to decay throughout train
ing. As a result, such methods are inapplicable to downlink (i.e., from the para
meter server to clients) compression in the cross-device setting, where heteroge
neous clients may appear only once during training and thus must download the mo
del parameters. Accordingly, we propose DoCoFL – a new framework for downlink co
mpression in the cross-device setting. Importantly, DoCoFL can be seamlessly com
bined with many uplink compression schemes, rendering it suitable for bi-directi
onal compression. Through extensive evaluation, we show that DoCoFL offers signi
ficant bi-directional bandwidth reduction while achieving competitive accuracy t
o that of a baseline without any compression.
****

Meta-Learning the Inductive Bias of Simple Neural Circuits
Will Dorrell, Maria Yuffa, Peter E. Latham
Training data is always finite, making it unclear how to generalise to unseen si
tuations. But, animals do generalise, wielding Occam’s razor to select a parsimo
nious explanation of their observations. How they do this is called their induct
ive bias, and it is implicitly built into the operation of animals’ neural circu
its. This relationship between an observed circuit and its inductive bias is a u
seful explanatory window for neuroscience, allowing design choices to be underst
ood normatively. However, it is generally very difficult to map circuit structur
e to inductive bias. Here, we present a neural network tool to bridge this gap.
The tool meta-learns the inductive bias by learning functions that a neural circ
uit finds easy to generalise, since easy-to-generalise functions are exactly tho
se the circuit chooses to explain incomplete data. In systems with analytically
known inductive bias, i.e. linear and kernel regression, our tool recovers it. G
enerally, we show it can flexibly extract inductive biases from supervised learn
ers, including spiking neural networks, and show how it could be applied to real
 animals. Finally, we use our tool to interpret recent connectomic data illustra
ting our intended use: understanding the role of circuit features through the re
sulting inductive bias.
****

Self-Repellent Random Walks on General Graphs - Achieving Minimal Sampling Varia

nce via Nonlinear Markov Chains
Vishwaraj Doshi, Jie Hu, Do Young Eun
We consider random walks on discrete state spaces, such as general undirected gr aphs, where the random walkers are designed to approximate a target quantity ove r the network topology via sampling and neighborhood exploration in the form of Markov chain Monte Carlo (MCMC) procedures. Given any Markov chain corresponding to a target probability distribution, we design a self-repellent random walk (S RRW) which is less likely to transition to nodes that were highly visited in the past, and more likely to transition to seldom visited nodes. For a class of SRR Ws parameterized by a positive real α, we prove that the empirical distri bution of the process converges almost surely to the the target (stationary) dis tribution of the underlying Markov chain kernel. We then provide a central limit theorem and derive the exact form of the arising asymptotic co-variance matrix, which allows us to show that the SRRW with a stronger repellence (larger $\alph a$) always achieves a smaller asymptotic covariance, in the sense of Loewner ord ering of co-variance matrices. Especially for SRRW-driven MCMC algorithms, we sh ow that the decrease in the asymptotic sampling variance is of the order $O(1/\a lpha)$, eventually going down to zero. Finally, we provide numerical simulations complimentary to our theoretical results, also empirically testing a version of SRRW with α increasing in time to combine the benefits of smaller asympt otic variance due to large α, with empirically observed faster mixing pro perties of SRRW with smaller α.
****
Linear Time GPs for Inferring Latent Trajectories from Neural Spike Trains
Matthew Dowling, Yuan Zhao, Il Memming Park
Latent Gaussian process (GP) models are widely used in neuroscience to uncover h idden state evolutions from sequential observations, mainly in neural activity r ecordings. While latent GP models provide a principled and powerful solution in theory, the intractable posterior in non-conjugate settings necessitates approxi mate inference schemes, which may lack scalability. In this work, we propose cvH M, a general inference framework for latent GP models leveraging Hida-Matérn ker nels and conjugate computation variational inference (CVI). With cvHM, we are ab le to perform variational inference of latent neural trajectories with linear ti me complexity for arbitrary likelihoods. The reparameterization of stationary ke rnels using Hida-Matérn GPs helps us connect the latent variable models that enc ode prior assumptions through dynamical systems to those that encode trajectory assumptions through GPs. In contrast to previous work, we use bidirectional info rmation filtering, leading to a more concise implementation. Furthermore, we emp loy the Whittle approximate likelihood to achieve highly efficient hyperparamete r learning.
****
On the Convergence Rate of Gaussianization with Random Rotations
Felix Draxler, Lars Kühmichel, Armand Rousselot, Jens Müller, Christoph Schnoerr , Ullrich Koethe
Gaussianization is a simple generative model that can be trained without backpro pagation. It has shown compelling performance on low dimensional data. As the di mension increases, however, it has been observed that the convergence speed slow s down. We show analytically that the number of required layers scales linearly with the dimension for Gaussian input. We argue that this is because the model i s unable to capture dependencies between dimensions. Empirically, we find the sa me linear increase in cost for arbitrary input $p(x)$, but observe favorable sca ling for some distributions. We explore potential speed-ups and formulate challe nges for further research.
****
PaLM-E: An Embodied Multimodal Language Model
Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Br ian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang , Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Van houcke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, Pe te Florence

Large language models excel at a wide range of complex tasks. However, enabling general inference in the real world, e.g. for robotics problems, raises the challenge of grounding. We propose embodied language models to directly incorporate real-world continuous sensor modalities into language models and thereby establish the link between words and percepts. Input to our embodied language model are multimodal sentences that interleave visual, continuous state estimation, and textual input encodings. We train these encodings end-to-end, in conjunction with a pre-trained large language model, for multiple embodied tasks including sequential robotic manipulation planning, visual question answering, and captioning. Our evaluations show that PaLM-E, a single large embodied multimodal model, can address a variety of embodied reasoning tasks, from a variety of observation modalities, on multiple embodiments, and further, exhibits positive transfer: the model benefits from diverse joint training across internet-scale language, vision, and visual-language domains. Our largest model with 562B parameters, in addition to being trained on robotics tasks, is a visual-language generalist with state-of-the-art performance on OK-VQA, and retains generalist language capabilities with increasing scale.

****

Reduce, Reuse, Recycle: Compositional Generation with Energy-Based Diffusion Models and MCMC

Yilun Du, Conor Durkan, Robin Strudel, Joshua B. Tenenbaum, Sander Dieleman, Rob Fergus, Jascha Sohl-Dickstein, Arnaud Doucet, Will Sussman Grathwohl

Since their introduction, diffusion models have quickly become the prevailing approach to generative modeling in many domains. They can be interpreted as learning the gradients of a time-varying sequence of log-probability density functions. This interpretation has motivated classifier-based and classifier-free guidance as methods for post-hoc control of diffusion models. In this work, we build upon these ideas using the score-based interpretation of diffusion models, and explore alternative ways to condition, modify, and reuse diffusion models for tasks involving compositional generation and guidance. In particular, we investigate why certain types of composition fail using current techniques and present a number of solutions. We conclude that the sampler (not the model) is responsible for this failure and propose new samplers, inspired by MCMC, which enable successful compositional generation. Further, we propose an energy-based parameterization of diffusion models which enables the use of new compositional operators and more sophisticated, Metropolis-corrected samplers. Intriguingly we find these samplers lead to notable improvements in compositional generation across a wide variety of problems such as classifier-guided ImageNet modeling and compositional text-to-image generation.

****

Multi-task Representation Learning for Pure Exploration in Linear Bandits

Yihan Du, Longbo Huang, Wen Sun

Despite the recent success of representation learning in sequential decision making, the study of the pure exploration scenario (i.e., identify the best option and minimize the sample complexity) is still limited. In this paper, we study multi-task representation learning for best arm identification in linear bandit (RepBAI-LB) and best policy identification in contextual linear bandit (RepBPI-CLB), two popular pure exploration settings with wide applications, e.g., clinical trials and web content optimization. In these two problems, all tasks share a common low-dimensional linear representation, and our goal is to leverage this feature to accelerate the best arm (policy) identification process for all tasks. For these problems, we design computationally and sample efficient algorithms DouExpDes and C-DouExpDes, which perform double experimental designs to plan optimal sample allocations for learning the global representation. We show that by learning the common representation among tasks, our sample complexity is significantly better than that of the native approach which solves tasks independently. To the best of our knowledge, this is the first work to demonstrate the benefits of representation learning for multi-task pure exploration.

****

Nonparametric Generative Modeling with Conditional Sliced-Wasserstein Flows

Chao Du, Tianbo Li, Tianyu Pang, Shuicheng Yan, Min Lin
Sliced-Wasserstein Flow (SWF) is a promising approach to nonparametric generativ e modeling but has not been widely adopted due to its suboptimal generative qual ity and lack of conditional modeling capabilities. In this work, we make two maj or contributions to bridging this gap. First, based on a pleasant observation th at (under certain conditions) the SWF of joint distributions coincides with thos e of conditional distributions, we propose Conditional Sliced-Wasserstein Flow (CSWF), a simple yet effective extension of SWF that enables nonparametric condit ional modeling. Second, we introduce appropriate inductive biases of images into SWF with two techniques inspired by local connectivity and multiscale represent ation in vision research, which greatly improve the efficiency and quality of mo deling images. With all the improvements, we achieve generative performance comp arable with many deep parametric generative models on both conditional and uncon ditional tasks in a purely nonparametric fashion, demonstrating its great potent ial.
****

Subsample Ridge Ensembles: Equivalences and Generalized Cross-Validation
Jin-Hong Du, Pratik Patil, Arun K. Kuchibhotla
We study subsampling-based ridge ensembles in the proportional asymptotics regim e, where the feature size grows proportionally with the sample size such that th eir ratio converges to a constant. By analyzing the squared prediction risk of r idge ensembles as a function of the explicit penalty λ and the limiting subsample aspect ratio ϕ_s (the ratio of the feature size to the subsample size), we characterize contours in the (λ, ϕ_s)-plane at any achievab le risk. As a consequence, we prove that the risk of the optimal full ridgeless ensemble (fitted on all possible subsamples) matches that of the optimal ridge p redictor. In addition, we prove strong uniform consistency of generalized cross- validation (GCV) over the subsample sizes for estimating the prediction risk of ridge ensembles. This allows for GCV-based tuning of full ridgeless ensembles wi thout sample splitting and yields a predictor whose risk matches optimal ridge r isk.
****

On Uni-Modal Feature Learning in Supervised Multi-Modal Learning
Chenzhuang Du, Jiaye Teng, Tingle Li, Yichen Liu, Tianyuan Yuan, Yue Wang, Yang Yuan, Hang Zhao
We abstract the features (i.e. learned representations) of multi-modal data into 1) uni-modal features, which can be learned from uni-modal training, and 2) pai red features, which can only be learned from cross-modal interactions. Multi-mod al models are expected to benefit from cross-modal interactions on the basis of ensuring uni-modal feature learning. However, recent supervised multi-modal late -fusion training approaches still suffer from insufficient learning of uni-modal features on each modality. We prove that this phenomenon does hurt the model's generalization ability. To this end, we propose to choose a targeted late-fusion learning method for the given supervised multi-modal task from Uni-Modal Ensemb le (UME) and the proposed Uni-Modal Teacher (UMT), according to the distribution of uni-modal and paired features. We demonstrate that, under a simple guiding s trategy, we can achieve comparable results to other complex late-fusion or inter mediate-fusion methods on various multi-modal datasets, including VGG-Sound, Kin etics-400, UCF101, and ModelNet40.
****

Guiding Pretraining in Reinforcement Learning with Large Language Models
Yuqing Du, Olivia Watkins, Zihan Wang, Cédric Colas, Trevor Darrell, Pieter Abbe el, Abhishek Gupta, Jacob Andreas
Reinforcement learning algorithms typically struggle in the absence of a dense, well-shaped reward function. Intrinsically motivated exploration methods address this limitation by rewarding agents for visiting novel states or transitions, b ut these methods offer limited benefits in large environments where most discove red novelty is irrelevant for downstream tasks. We describe a method that uses b ackground knowledge from text corpora to shape exploration. This method, called ELLM (Exploring with LLMs) rewards an agent for achieving goals suggested by a l

anguage model prompted with a description of the agent’s current state. By lever aging large-scale language model pretraining, ELLM guides agents toward human-me aningful and plausibly useful behaviors without requiring a human in the loop. W e evaluate ELLM in the Crafter game environment and the Housekeep robotic simula tor, showing that ELLM-trained agents have better coverage of common-sense behav iors during pretraining and usually match or improve performance on a range of d ownstream tasks.

****

A Flexible Diffusion Model

Weitao Du, He Zhang, Tao Yang, Yuanqi Du

Denoising diffusion (score-based) generative models have become a popular choice for modeling complex data. Recently, a deep connection between forward-backward stochastic differential equations (SDEs) and diffusion-based models has been es tablished, leading to the development of new SDE variants such as sub-VP and cri tically-damped Langevin. Despite the empirical success of some hand-crafted forw ard SDEs, many potentially promising forward SDEs remain unexplored. In this wor k, we propose a general framework for parameterizing diffusion models, particula rly the spatial part of forward SDEs, by leveraging the symplectic and Riemannia n geometry of the data manifold. We introduce a systematic formalism with theore tical guarantees and connect it with previous diffusion models. Finally, we demo nstrate the theoretical advantages of our method from a variational optimization perspective. We present numerical experiments on synthetic datasets, MNIST and CIFAR10 to validate the effectiveness of our framework.

****

Fast Excess Risk Rates via Offset Rademacher Complexity

Chenguang Duan, Yuling Jiao, Lican Kang, Xiliang Lu, Jerry Zhijian Yang

Based on the offset Rademacher complexity, this work outlines a systematical fra mework for deriving sharp excess risk bounds in statistical learning without Ber nstein condition. In addition to recovering fast rates in a unified way for some parametric and nonparametric supervised learning models with minimum identifiab ility assumptions, we also obtain new and improved results for LAD (sparse) line ar regression and deep logistic regression with deep ReLU neural networks, respe ctively.

****

Are Diffusion Models Vulnerable to Membership Inference Attacks?

Jinhao Duan, Fei Kong, Shiqi Wang, Xiaoshuang Shi, Kaidi Xu

Diffusion-based generative models have shown great potential for image synthesis , but there is a lack of research on the security and privacy risks they may pos e. In this paper, we investigate the vulnerability of diffusion models to Member ship Inference Attacks (MIAs), a common privacy concern. Our results indicate th at existing MIAs designed for GANs or VAE are largely ineffective on diffusion m odels, either due to inapplicable scenarios (e.g., requiring the discriminator o f GANs) or inappropriate assumptions (e.g., closer distances between synthetic s amples and member samples). To address this gap, we propose Step-wise Error Comp aring Membership Inference (SecMI), a query-based MIA that infers memberships by assessing the matching of forward process posterior estimation at each timestep . SecMI follows the common overfitting assumption in MIA where member samples no rmally have smaller estimation errors, compared with hold-out samples. We consid er both the standard diffusion models, e.g., DDPM, and the text-to-image diffusi on models, e.g., Latent Diffusion Models and Stable Diffusion. Experimental resu lts demonstrate that our methods precisely infer the membership with high confid ence on both of the two scenarios across multiple different datasets. Code is av ailable at https://github.com/jinhaoduan/SecMI.

****

Bayesian Progressive Deep Topic Model with Knowledge Informed Textual Data Coars ening Process

Zhibin Duan, Xinyang Liu, Yudi Su, Yishi Xu, Bo Chen, Mingyuan Zhou

Deep topic models have shown an impressive ability to extract multi-layer docume nt latent representations and discover hierarchical semantically meaningful topi cs.However, most deep topic models are limited to the single-step generative pro

cess, despite the fact that the progressive generative process has achieved impr
essive performance in modeling image data. To this end, in this paper, we propos
e a novel progressive deep topic model that consists of a knowledge-informed tex
tural data coarsening process and a corresponding progressive generative model.
The former is used to build multi-level observations ranging from concrete to ab
stract, while the latter is used to generate more concrete observations graduall
y. Additionally, we incorporate a graph-enhanced decoder to capture the semantic
 relationships among words at different levels of observation. Furthermore, we p
erform a theoretical analysis of the proposed model based on the principle of in
formation theory and show how it can alleviate the well-known "latent variable c
ollapse" problem. Finally, extensive experiments demonstrate that our proposed m
odel effectively improves the ability of deep topic models, resulting in higher-
quality latent document representations and topics.
****

Are Equivariant Equilibrium Approximators Beneficial?
Zhijian Duan, Yunxuan Ma, Xiaotie Deng
Recently, remarkable progress has been made by approximating Nash equilibrium (N
E), correlated equilibrium (CE), and coarse correlated equilibrium (CCE) through
 function approximation that trains a neural network to predict equilibria from
game representations. Furthermore, equivariant architectures are widely adopted
in designing such equilibrium approximators in normal-form games. In this paper,
 we theoretically characterize the benefits and limitations of equivariant equil
ibrium approximators. For the benefits, we show that they enjoy better generaliz
ability than general ones and can achieve better approximations when the payoff
distribution is permutation-invariant. For the limitations, we discuss their dra
wbacks in terms of equilibrium selection and social welfare. Together, our resul
ts help to understand the role of equivariance in equilibrium approximators.
****

Evaluating Self-Supervised Learning via Risk Decomposition
Yann Dubois, Tatsunori Hashimoto, Percy Liang
Self-supervised learning (SSL) is typically evaluated using a single metric (lin
ear probing on ImageNet), which neither provides insight into tradeoffs between
models nor highlights how to improve them. To address this, we propose an SSL ri
sk decomposition, which generalizes the classical approximation-estimation decom
position. Our decomposition consists of four error terms: approximation, represe
ntation usability, probe generalization, and encoder generalization. We provide
efficient estimators for each term and use them to analyze the effect of 30 desi
gn choices on 169 SSL vision models evaluated on ImageNet. Our analysis gives va
luable insights for designing and using SSL models. For example, it highlights t
he main source of errors and shows how to improve SSL in specific settings (full
- vs few-shot) by trading off error components.
****

Fully Dynamic Submodular Maximization over Matroids
Paul Duetting, Federico Fusco, Silvio Lattanzi, Ashkan Norouzi-Fard, Morteza Zad
imoghaddam
Maximizing monotone submodular functions under a matroid constraint is a classic
 algorithmic problem with multiple applications in data mining and machine learn
ing. We study this classic problem in the fully dynamic setting, where elements
can be both inserted and deleted in real-time. Our main result is a randomized a
lgorithm that maintains an efficient data structure with an $\tilde{O}(k^2)$ amo
rtized update time (in the number of additions and deletions) and yields a 4-a
pproximate solution, where k is the rank of the matroid.
****

Optimal No-Regret Learning for One-Sided Lipschitz Functions
Paul Duetting, Guru Guruganesh, Jon Schneider, Joshua Ruizhi Wang
Inspired by applications in pricing and contract design, we study the maximizati
on of one-sided Lipschitz functions, which only provide the (weaker) guarantee t
hat they do not grow too quickly in one direction. We show that it is possible t
o learn a maximizer for such a function while incurring $O(\log \log T)$ total r
egret (with a universal constant independent of the number of discontinuities /

complexity of the function). This regret bound is asymptotically optimal in T due to a lower bound of Kleinberg and Leighton. By applying this algorithm, we s how that one can sell digital goods to multiple buyers and learn the optimal lin ear contract in the principal-agent setting while incurring at most $O(\log \log T)$ regret.

****

Integrating Prior Knowledge in Contrastive Learning with Kernel

Benoit Dufumier, Carlo Alberto Barbano, Robin Louiset, Edouard Duchesnay, Pietro Gori

Data augmentation is a crucial component in unsupervised contrastive learning (C L). It determines how positive samples are defined and, ultimately, the quality of the learned representation. In this work, we open the door to new perspective s for CL by integrating prior knowledge, given either by generative models - vie wed as prior representations - or weak attributes in the positive and negative s ampling. To this end, we use kernel theory to propose a novel loss, called decou pled uniformity, that i) allows the integration of prior knowledge and ii) remov es the positive-negative coupling in the original InfoNCE loss. We draw a connec tion between contrastive learning and the conditional mean embedding theory to d erive tight bounds on the downstream classification loss. In an unsupervised set ting, we empirically demonstrate that CL benefits from generative models to impr ove its representation both on natural and medical images. In a weakly supervise d scenario, our framework outperforms other unconditional and conditional CL app roaches.

****

Q-Flow: Generative Modeling for Differential Equations of Open Quantum Dynamics with Normalizing Flows

Owen M Dugan, Peter Y. Lu, Rumen Dangovski, Di Luo, Marin Soljacic

Studying the dynamics of open quantum systems can enable breakthroughs both in f undamental physics and applications to quantum engineering and quantum computati on. Since the density matrix ρ, which is the fundamental description for th e dynamics of such systems, is high-dimensional, customized deep generative neur al networks have been instrumental in modeling ρ. However, the complex-valu ed nature and normalization constraints of ρ, as well as its complicated dy namics, prohibit a seamless connection between open quantum systems and the rece nt advances in deep generative modeling. Here we lift that limitation by utilizi ng a reformulation of open quantum system dynamics to a partial differential equ ation (PDE) for a corresponding probability distribution Q, the Husimi Q funct ion. Thus, we model the Q function seamlessly with off-the-shelf deep generative models such as normalizing flows. Additionally, we develop novel methods for le arning normalizing flow evolution governed by high-dimensional PDEs based on the Euler method and the application of the time-dependent variational principle. W e name the resulting approach Q-Flow and demonstrate the scalability and efficie ncy of Q-Flow on open quantum system simulations, including the dissipative harm onic oscillator and the dissipative bosonic model. Q-Flow is superior to convent ional PDE solvers and state-of-the-art physics-informed neural network solvers, especially in high-dimensional systems.

****

Adaptive Whitening in Neural Populations with Gain-modulating Interneurons

Lyndon Duong, David Lipshutz, David Heeger, Dmitri Chklovskii, Eero P Simoncelli

Statistical whitening transformations play a fundamental role in many computatio nal systems, and may also play an important role in biological sensory systems. Existing neural circuit models of adaptive whitening operate by modifying synapt ic interactions; however, such modifications would seem both too slow and insuff iciently reversible. Motivated by the extensive neuroscience literature on gain modulation, we propose an alternative model that adaptively whitens its response s by modulating the gains of individual neurons. Starting from a novel whitening objective, we derive an online algorithm that whitens its outputs by adjusting the marginal variances of an overcomplete set of projections. We map the algorit hm onto a recurrent neural network with fixed synaptic weights and gain-modulati ng interneurons. We demonstrate numerically that sign-constraining the gains imp

roves robustness of the network to ill-conditioned inputs, and a generalization of the circuit achieves a form of local whitening in convolutional populations, such as those found throughout the visual or auditory systems.
****

Generalization Bounds using Data-Dependent Fractal Dimensions
Benjamin Dupuis, George Deligiannidis, Umut Simsekli
Providing generalization guarantees for modern neural networks has been a crucial task in statistical learning. Recently, several studies have attempted to analyze the generalization error in such settings by using tools from fractal geometry. While these works have successfully introduced new mathematical tools to apprehend generalization, they heavily rely on a Lipschitz continuity assumption, which in general does not hold for neural networks and might make the bounds vacuous. In this work, we address this issue and prove fractal geometry-based generalization bounds without requiring any Lipschitz assumption. To achieve this goal, we build up on a classical covering argument in learning theory and introduce a data-dependent fractal dimension. Despite introducing a significant amount of technical complications, this new notion lets us control the generalization error (over either fixed or random hypothesis spaces) along with certain mutual information (MI) terms. To provide a clearer interpretation to the newly introduced MI terms, as a next step, we introduce a notion of ‘geometric stability’ and link our bounds to the prior art. Finally, we make a rigorous connection between the proposed data-dependent dimension and topological data analysis tools, which then enables us to compute the dimension in a numerically efficient way. We support our theory with experiments conducted on various settings.
****

Multi-Objective Population Based Training
Arkadiy Dushatskiy, Alexander Chebykin, Tanja Alderliesten, Peter Bosman
Population Based Training (PBT) is an efficient hyperparameter optimization algorithm. PBT is a single-objective algorithm, but many real-world hyperparameter optimization problems involve two or more conflicting objectives. In this work, we therefore introduce a multi-objective version of PBT, MO-PBT. Our experiments on diverse multi-objective hyperparameter optimization problems (Precision/Recall, Accuracy/Fairness, Accuracy/Adversarial Robustness) show that MO-PBT outperforms random search, single-objective PBT, and the state-of-the-art multi-objective hyperparameter optimization algorithm MO-ASHA.
****

Neural Diffusion Processes
Vincent Dutordoir, Alan Saul, Zoubin Ghahramani, Fergus Simpson
Neural network approaches for meta-learning distributions over functions have desirable properties such as increased flexibility and a reduced complexity of inference. Building on the successes of denoising diffusion models for generative modelling, we propose Neural Diffusion Processes (NDPs), a novel approach that learns to sample from a rich distribution over functions through its finite marginals. By introducing a custom attention block we are able to incorporate properties of stochastic processes, such as exchangeability, directly into the NDP’s architecture. We empirically show that NDPs can capture functional distributions close to the true Bayesian posterior, demonstrating that they can successfully emulate the behaviour of Gaussian processes and surpass the performance of neural processes. NDPs enable a variety of downstream tasks, including regression, implicit hyperparameter marginalisation, non-Gaussian posterior prediction and global optimisation.
****

FAENet: Frame Averaging Equivariant GNN for Materials Modeling
Alexandre Agm Duval, Victor Schmidt, Alex Hernández-Garc■■a, Santiago Miret, Fragkiskos D. Malliaros, Yoshua Bengio, David Rolnick
Applications of machine learning techniques for materials modeling typically involve functions that are known to be equivariant or invariant to specific symmetries. While graph neural networks (GNNs) have proven successful in such applications, conventional GNN approaches that enforce symmetries via the model architecture often reduce expressivity, scalability or comprehensibility. In this paper,

we introduce (1) a flexible, model-agnostic framework based on stochastic frame averaging that enforces E(3) equivariance or invariance, without any architectural constraints; (2) FAENet: a simple, fast and expressive GNN that leverages stochastic frame averaging to process geometric information without constraints. We prove the validity of our method theoretically and demonstrate its superior accuracy and computational scalability in materials modeling on the OC20 dataset (S2EF, IS2RE) as well as common molecular modeling tasks (QM9, QM7-X).
****
Blackout Diffusion: Generative Diffusion Models in Discrete-State Spaces
Javier E. Santos, Zachary R. Fox, Nicholas Lubbers, Yen Ting Lin
Typical generative diffusion models rely on a Gaussian diffusion process for training the backward transformations, which can then be used to generate samples from Gaussian noise. However, real world data often takes place in discrete-state spaces, including many scientific applications. Here, we develop a theoretical formulation for arbitrary discrete-state Markov processes in the forward diffusion process using exact (as opposed to variational) analysis. We relate the theory to the existing continuous-state Gaussian diffusion as well as other approaches to discrete diffusion, and identify the corresponding reverse-time stochastic process and score function in the continuous-time setting, and the reverse-time mapping in the discrete-time setting. As an example of this framework, we introduce “Blackout Diffusion”, which learns to produce samples from an empty image instead of from noise. Numerical experiments on the CIFAR-10, Binarized MNIST, and CelebA datasets confirm the feasibility of our approach. Generalizing from specific (Gaussian) forward processes to discrete-state processes without a variational approximation sheds light on how to interpret diffusion models, which we discuss.
****
The Computational Complexity of Concise Hypersphere Classification
Eduard Eiben, Robert Ganian, Iyad A. Kanj, Sebastian Ordyniak, Stefan Szeider
Hypersphere classification is a classical and foundational method that can provide easy-to-process explanations for the classification of real-valued as well as binary data. However, obtaining an (ideally concise) explanation via hypersphere classification is much more difficult when dealing with binary data as opposed to real-valued data. In this paper, we perform the first complexity-theoretic study of the hypersphere classification problem for binary data. We use the fine-grained parameterized complexity paradigm to analyze the impact of structural properties that may be present in the input data as well as potential conciseness constraints. Our results include not only stronger lower bounds but also a number of new fixed-parameter algorithms for hypersphere classification of binary data, which can find an exact and concise explanation when one exists.
****
E(n) Equivariant Message Passing Simplicial Networks
Floor Eijkelboom, Rob Hesselink, Erik J Bekkers
This paper presents $\mathrm{E}(n)$ Equivariant Message Passing Simplicial Networks (EMPSNs), a novel approach to learning on geometric graphs and point clouds that is equivariant to rotations, translations, and reflections. EMPSNs can learn high-dimensional simplex features in graphs (e.g. triangles), and use the increase of geometric information of higher-dimensional simplices in an $\mathrm{E}(n)$ equivariant fashion. EMPSNs simultaneously generalize $\mathrm{E}(n)$ Equivariant Graph Neural Networks to a topologically more elaborate counterpart and provide an approach for including geometric information in Message Passing Simplicial Networks, thereby serving as a proof of concept for combining geometric and topological information in graph learning. The results indicate that EMPSNs can leverage the benefits of both approaches, leading to a general increase in performance when compared to either method individually, being on par with state-of-the-art approaches for learning on geometric graphs. Moreover, the results suggest that incorporating geometric information serves as an effective measure against over-smoothing in message passing networks, especially when operating on high-dimensional simplicial structures.
****

Performative Recommendation: Diversifying Content via Strategic Incentives
Itay Eilat, Nir Rosenfeld
The primary goal in recommendation is to suggest relevant content to users, but optimizing for accuracy often results in recommendations that lack diversity. To remedy this, conventional approaches such as re-ranking improve diversity by presenting more diverse items. Here we argue that to promote inherent and prolonged diversity, the system must encourage its creation. Towards this, we harness the performative nature of recommendation, and show how learning can incentivize strategic content creators to create diverse content. Our approach relies on a novel form of regularization that anticipates strategic changes to content, and penalizes for content homogeneity. We provide analytic and empirical results that demonstrate when and how diversity can be incentivized, and experimentally demonstrate the utility of our approach on synthetic and semi-synthetic data.
****
Hyperparameters in Reinforcement Learning and How To Tune Them
Theresa Eimer, Marius Lindauer, Roberta Raileanu
In order to improve reproducibility, deep reinforcement learning (RL) has been adopting better scientific practices such as standardized evaluation metrics and reporting. However, the process of hyperparameter optimization still varies widely across papers, which makes it challenging to compare RL algorithms fairly. In this paper, we show that hyperparameter choices in RL can significantly affect the agent’s final performance and sample efficiency, and that the hyperparameter landscape can strongly depend on the tuning seed which may lead to overfitting. We therefore propose adopting established best practices from AutoML, such as the separation of tuning and testing seeds, as well as principled hyperparameter optimization (HPO) across a broad search space. We support this by comparing multiple state-of-the-art HPO tools on a range of RL algorithms and environments to their hand-tuned counterparts, demonstrating that HPO approaches often have higher performance and lower compute overhead. As a result of our findings, we recommend a set of best practices for the RL community, which should result in stronger empirical results with fewer computational costs, better reproducibility, and thus faster progress. In order to encourage the adoption of these practices, we provide plug-and-play implementations of the tuning algorithms used in this paper at https://github.com/facebookresearch/how-to-autorl.
****
Fairness in Streaming Submodular Maximization over a Matroid Constraint
Marwa El Halabi, Federico Fusco, Ashkan Norouzi-Fard, Jakab Tardos, Jakub Tarnawski
Streaming submodular maximization is a natural model for the task of selecting a representative subset from a large-scale dataset. If datapoints have sensitive attributes such as gender or race, it becomes important to enforce fairness to avoid bias and discrimination. This has spurred significant interest in developing fair machine learning algorithms. Recently, such algorithms have been developed for monotone submodular maximization under a cardinality constraint. In this paper, we study the natural generalization of this problem to a matroid constraint. We give streaming algorithms as well as impossibility results that provide trade-offs between efficiency, quality and fairness. We validate our findings empirically on a range of well-known real-world applications: exemplar-based clustering, movie recommendation, and maximum coverage in social networks.
****
Difference of submodular minimization via DC programming
Marwa El Halabi, George Orfanides, Tim Hoheisel
Minimizing the difference of two submodular (DS) functions is a problem that naturally occurs in various machine learning problems. Although it is well known that a DS problem can be equivalently formulated as the minimization of the difference of two convex (DC) functions, existing algorithms do not fully exploit this connection. A classical algorithm for DC problems is called the DC algorithm (DCA). We introduce variants of DCA and its complete form (CDCA) that we apply to the DC program corresponding to DS minimization. We extend existing convergence properties of DCA, and connect them to convergence properties on the DS problem.

Our results on DCA match the theoretical guarantees satisfied by existing DS al
gorithms, while providing a more complete characterization of convergence proper
ties. In the case of CDCA, we obtain a stronger local minimality guarantee. Our
numerical results show that our proposed algorithms outperform existing baseline
s on two applications: speech corpus selection and feature selection.
****
Graph Positional Encoding via Random Feature Propagation
Moshe Eliasof, Fabrizio Frasca, Beatrice Bevilacqua, Eran Treister, Gal Chechik,
Haggai Maron
Two main families of node feature augmentation schemes have been explored for en
hancing GNNs: random features and spectral positional encoding. Surprisingly, ho
wever, there is still no clear understanding of the relation between these two a
ugmentation schemes. Here we propose a novel family of positional encoding schem
es which draws a link between the above two approaches and improves over both. T
he new approach, named Random Feature Propagation (RFP), is inspired by the powe
r iteration method and its generalizations. It concatenates several intermediate
steps of an iterative algorithm for computing the dominant eigenvectors of a pr
opagation matrix, starting from random node features. Notably, these propagation
steps are based on graph-dependent propagation operators that can be either pre
defined or learned. We explore the theoretical and empirical benefits of RFP. Fi
rst, we provide theoretical justifications for using random features, for incorp
orating early propagation steps, and for using multiple random initializations.
Then, we empirically demonstrate that RFP significantly outperforms both spectra
l PE and random features in multiple node classification and graph classificatio
n benchmarks.
****
Improving Graph Neural Networks with Learnable Propagation Operators
Moshe Eliasof, Lars Ruthotto, Eran Treister
Graph Neural Networks (GNNs) are limited in their propagation operators. In many
cases, these operators often contain non-negative elements only and are shared
across channels, limiting the expressiveness of GNNs. Moreover, some GNNs suffer
from over-smoothing, limiting their depth. On the other hand, Convolutional Neu
ral Networks (CNNs) can learn diverse propagation filters, and phenomena like ov
er-smoothing are typically not apparent in CNNs. In this paper, we bridge these
gaps by incorporating trainable channel-wise weighting factors ω to learn
and mix multiple smoothing and sharpening propagation operators at each layer.
Our generic method is called ωGNN, and is easy to implement. We study two
variants: ωGCN and ωGAT. For ωGCN, we theoretically analys
e its behaviour and the impact of ω on the obtained node features. Our ex
periments confirm these findings, demonstrating and explaining how both variants
do not over-smooth. Additionally, we experiment with 15 real-world datasets on
node- and graph-classification tasks, where our ωGCN and ωGAT perf
orm on par with state-of-the-art methods.
****
Phase Transitions in the Detection of Correlated Databases
Dor Elimelech, Wasim Huleihel
We study the problem of detecting the correlation between two Gaussian databases
$\mathsf{X}\in\mathbb{R}^{n\times d}$ and $\mathsf{Y}^{n\times d}$, each compos
ed of n users with d features. This problem is relevant in the analysis of s
ocial media, computational biology, etc. We formulate this as a hypothesis testi
ng problem: under the null hypothesis, these two databases are statistically ind
ependent. Under the alternative, however, there exists an unknown permutation $\
sigma$ over the set of n users (or, row permutation), such that X i
s ρ-correlated with Y^σ, a permuted version of Y
. We determine sharp thresholds at which optimal testing exhibits a phase transi
tion, depending on the asymptotic regime of n and d. Specifically, we prove
that if $\rho^2d\to0$, as $d\to\infty$, then weak detection (performing slightly
better than random guessing) is statistically impossible, irrespectively of the
value of n. This compliments the performance of a simple test that thresholds
the sum all entries of $\mathsf{X}^T\mathsf{Y}$. Furthermore, when d is fixed

, we prove that strong detection (vanishing error probability) is impossible for any $\rho<\rho^\star$, where $\rho^\star$ is an explicit function of d, while weak detection is again impossible as long as $\rho^2d=o(1)$, as $n\to\infty$. These results close significant gaps in current recent related studies.
****
A new near-linear time algorithm for k-nearest neighbor search using a compressed cover tree
Yury Elkin, Vitaliy Kurlin
Given a reference set R of n points and a query set Q of m points in a metric space, this paper studies an important problem of finding k-nearest neighbors of every point q of Q in the set R in a near-linear time. In the paper at ICML 2006, Beygelzimer, Kakade, and Langford introduced a cover tree and attempted to prove that this tree can be built in O(n log n) time while the nearest neighbor search can be done O(n log m) time with a hidden dimensionality factor. In 2015, section 5.3 of Curtin's PhD pointed out that the proof of the latter claim can have a serious gap in time complexity estimation. A paper at TopoInVis 2022 reported explicit counterexamples for a key step in the proofs of both claims. The past obstacles will be overcome by a simpler compressed cover tree on the reference set R. The first new algorithm constructs a compressed cover tree in O(n log n) time. The second new algorithm finds all k-nearest neighbors of all points from Q using a compressed cover tree in time O(m(k+log n)log k) with a hidden dimensionality factor depending on point distributions of the sets R,Q but not on their sizes.
****
Motion Question Answering via Modular Motion Programs
Mark Endo, Joy Hsu, Jiaman Li, Jiajun Wu
In order to build artificial intelligence systems that can perceive and reason with human behavior in the real world, we must first design models that conduct complex spatio-temporal reasoning over motion sequences. Moving towards this goal, we propose the HumanMotionQA task to evaluate complex, multi-step reasoning abilities of models on long-form human motion sequences. We generate a dataset of question-answer pairs that require detecting motor cues in small portions of motion sequences, reasoning temporally about when events occur, and querying specific motion attributes. In addition, we propose NSPose, a neuro-symbolic method for this task that uses symbolic reasoning and a modular design to ground motion through learning motion concepts, attribute neural operators, and temporal relations. We demonstrate the suitability of NSPose for the HumanMotionQA task, outperforming all baseline methods.
****
Learning Perturbations to Explain Time Series Predictions
Joseph Enguehard
Explaining predictions based on multivariate time series data carries the additional difficulty of handling not only multiple features, but also time dependencies. It matters not only what happened, but also when, and the same feature could have a very different impact on a prediction depending on this time information. Previous work has used perturbation-based saliency methods to tackle this issue, perturbing an input using a trainable mask to discover which features at which times are driving the predictions. However these methods introduce fixed perturbations, inspired from similar methods on static data, while there seems to be little motivation to do so on temporal data. In this work, we aim to explain predictions by learning not only masks, but also associated perturbations. We empirically show that learning these perturbations significantly improves the quality of these explanations on time series data.
****
Regret Minimization and Convergence to Equilibria in General-sum Markov Games
Liad Erez, Tal Lancewicki, Uri Sherman, Tomer Koren, Yishay Mansour
An abundance of recent impossibility results establish that regret minimization in Markov games with adversarial opponents is both statistically and computationally intractable. Nevertheless, none of these results preclude the possibility of regret minimization under the assumption that all parties adopt the same learn

ing procedure. In this work, we present the first (to our knowledge) algorithm f or learning in general-sum Markov games that provides sublinear regret guarantee s when executed by all agents. The bounds we obtain are for $\textit{swap regret }$, and thus, along the way, imply convergence to a $\textit{correlated}$ equili brium. Our algorithm is decentralized, computationally efficient, and does not r equire any communication between agents. Our key observation is that online lear ning via policy optimization in Markov games essentially reduces to a form of $\ textit{weighted}$ regret minimization, with $\textit{unknown}$ weights determine d by the path length of the agents' policy sequence. Consequently, controlling t he path length leads to weighted regret objectives for which sufficiently adapti ve algorithms provide sublinear regret guarantees.

****

Delayed Bandits: When Do Intermediate Observations Help?

Emmanuel Esposito, Saeed Masoudian, Hao Qiu, Dirk Van Der Hoeven, Nicolò Cesa-Bi anchi, Yevgeny Seldin

We study a K-armed bandit with delayed feedback and intermediate observations. We consider a model, where intermediate observations have a form of a finite st ate, which is observed immediately after taking an action, whereas the loss is o bserved after an adversarially chosen delay. We show that the regime of the mapp ing of states to losses determines the complexity of the problem, irrespective o f whether the mapping of actions to states is stochastic or adversarial. If the mapping of states to losses is adversarial, then the regret rate is of order $\s qrt{(K+d)T}$ (within log factors), where T is the time horizon and d is a fi xed delay. This matches the regret rate of a K-armed bandit with delayed feedb ack and without intermediate observations, implying that intermediate observatio ns are not helpful. However, if the mapping of states to losses is stochastic, w e show that the regret grows at a rate of $\sqrt{\bigl(K+\min\{|\mathcal{S}|,d\} \bigr)T}$ (within log factors), implying that if the number $|\mathcal{S}|$ of s tates is smaller than the delay, then intermediate observations help. We also pr ovide refined high-probability regret upper bounds for non-uniform delays, toget her with experimental validation of our algorithms.

****

Scaling Spherical CNNs

Carlos Esteves, Jean-Jacques Slotine, Ameesh Makadia

Spherical CNNs generalize CNNs to functions on the sphere, by using spherical co nvolutions as the main linear operation. The most accurate and efficient way to compute spherical convolutions is in the spectral domain (via the convolution th eorem), which is still costlier than the usual planar convolutions. For this rea son, applications of spherical CNNs have so far been limited to small problems t hat can be approached with low model capacity. In this work, we show how spheric al CNNs can be scaled for much larger problems. To achieve this, we make critica l improvements including novel variants of common model components, an implement ation of core operations to exploit hardware accelerator characteristics, and ap plication-specific input representations that exploit the properties of our mode l. Experiments show our larger spherical CNNs reach state-of-the-art on several targets of the QM9 molecular benchmark, which was previously dominated by equiva riant graph neural networks, and achieve competitive performance on multiple wea ther forecasting tasks. Our code is available at https://github.com/google-resea rch/spherical-cnn.

****

Stochastic Gradient Descent under Markovian Sampling Schemes

Mathieu Even

We study a variation of vanilla stochastic gradient descent where the optimizer only has access to a Markovian sampling scheme. These schemes encompass applicat ions that range from decentralized optimization with a random walker (token algo rithms), to RL and online system identification problems. We focus on obtaining rates of convergence under the least restrictive assumptions possible on the und erlying Markov chain and on the functions optimized. We first unveil the theoret ical lower bound for methods that sample stochastic gradients along the path of a Markov chain, making appear a dependency in the hitting time of the underlying

Markov chain. We then study Markov chain SGD (MC-SGD) under much milder regularity assumptions than prior works. We finally introduce MC-SAG, an alternative to MC-SGD with variance reduction, that only depends on the hitting time of the Markov chain, therefore obtaining a communication-efficient token algorithm.

****

Continual Learning in Linear Classification on Separable Data

Itay Evron, Edward Moroshko, Gon Buzaglo, Maroun Khriesh, Badea Marjieh, Nathan Srebro, Daniel Soudry

We analyze continual learning on a sequence of separable linear classification tasks with binary labels. We show theoretically that learning with weak regularization reduces to solving a sequential max-margin problem, corresponding to a special case of the Projection Onto Convex Sets (POCS) framework. We then develop upper bounds on the forgetting and other quantities of interest under various settings with recurring tasks, including cyclic and random orderings of tasks. We discuss several practical implications to popular training practices like regularization scheduling and weighting. We point out several theoretical differences between our continual classification setting and a recently studied continual regression setting.

****

A Connection between One-Step RL and Critic Regularization in Reinforcement Learning

Benjamin Eysenbach, Matthieu Geist, Sergey Levine, Ruslan Salakhutdinov

As with any machine learning problem with limited data, effective offline RL algorithms require careful regularization to avoid overfitting. One class of methods, known as one-step RL, perform just one step of policy improvement. These methods, which include advantage-weighted regression and conditional behavioral cloning, are thus simple and stable, but can have limited asymptotic performance. A second class of methods, known as critic regularization, perform many steps of policy improvement with a regularized objective. These methods typically require more compute but have appealing lower-bound guarantees. In this paper, we draw a connection between these methods: applying a multi-step critic regularization method with a regularization coefficient of 1 yields the same policy as one-step RL. While our theoretical results require assumptions (e.g., deterministic dynamics), our experiments nevertheless show that our analysis makes accurate, testable predictions about practical offline RL methods (CQL and one-step RL) with commonly-used hyperparameters.

****

Neural Status Registers

Lukas Faber, Roger Wattenhofer

We study the problem of learning comparisons between numbers with neural networks. Despite comparisons being a seemingly simple problem, we find that both general-purpose models such as multilayer perceptrons (MLPs) as well as arithmetic architectures such as the Neural Arithmetic Logic Unit (NALU) struggle with learning comparisons. Neither architecture can extrapolate to much larger numbers than those seen in the training set. We propose a novel differentiable architecture, the Neural Status Register (NSR) to solve this problem. We experimentally validate the NSR in various settings. We can combine the NSR with other neural models to solve interesting problems such as piecewise-defined arithmetic, comparison of digit images, recurrent problems, or finding shortest paths in graphs. The NSR outperforms all baseline architectures, especially when it comes to extrapolating to larger numbers.

****

Learning Rate Schedules in the Presence of Distribution Shift

Matthew Fahrbach, Adel Javanmard, Vahab Mirrokni, Pratik Worah

We design learning rate schedules that minimize regret for SGD-based online learning in the presence of a changing data distribution. We fully characterize the optimal learning rate schedule for online linear regression via a novel analysis with stochastic differential equations. For general convex loss functions, we propose new learning rate schedules that are robust to distribution shift, and give upper and lower bounds for the regret that only differ by constants. For non-

convex loss functions, we define a notion of regret based on the gradient norm o
f the estimated models and propose a learning schedule that minimizes an upper b
ound on the total expected regret. Intuitively, one expects changing loss landsc
apes to require more exploration, and we confirm that optimal learning rate sche
dules typically have higher learning rates in the presence of distribution shift
. Finally, we provide experiments that illustrate these learning rate schedules
and their regret.
****
Predicting Rare Events by Shrinking Towards Proportional Odds
Gregory Faletto, Jacob Bien
Training classifiers is difficult with severe class imbalance, but many rare eve
nts are the culmination of a sequence with much more common intermediate outcome
s. For example, in online marketing a user first sees an ad, then may click on i
t, and finally may make a purchase; estimating the probability of purchases is d
ifficult because of their rarity. We show both theoretically and through data ex
periments that the more abundant data in earlier steps may be leveraged to impro
ve estimation of probabilities of rare events. We present PRESTO, a relaxation o
f the proportional odds model for ordinal regression. Instead of estimating weig
hts for one separating hyperplane that is shifted by separate intercepts for eac
h of the estimated Bayes decision boundaries between adjacent pairs of categoric
al responses, we estimate separate weights for each of these transitions. We imp
ose an L1 penalty on the differences between weights for the same feature in adj
acent weight vectors in order to shrink towards the proportional odds model. We
prove that PRESTO consistently estimates the decision boundary weights under a s
parsity assumption. Synthetic and real data experiments show that our method can
 estimate rare probabilities in this setting better than both logistic regressio
n on the rare category, which fails to borrow strength from more abundant catego
ries, and the proportional odds model, which is too inflexible.
****
Free-Form Variational Inference for Gaussian Process State-Space Models
Xuhui Fan, Edwin V. Bonilla, Terence O’Kane, Scott A Sisson
Gaussian process state-space models (GPSSMs) provide a principled and flexible a
pproach to modeling the dynamics of a latent state, which is observed at discret
e-time points via a likelihood model. However, inference in GPSSMs is computatio
nally and statistically challenging due to the large number of latent variables
in the model and the strong temporal dependencies between them. In this paper, w
e propose a new method for inference in Bayesian GPSSMs, which overcomes the dra
wbacks of previous approaches, namely over-simplified assumptions, and high comp
utational requirements. Our method is based on free-form variational inference v
ia stochastic gradient Hamiltonian Monte Carlo within the inducing-variable form
alism. Furthermore, by exploiting our proposed variational distribution, we prov
ide a collapsed extension of our method where the inducing variables are margina
lized analytically. We also showcase results when combining our framework with p
article MCMC methods. We show that, on six real-world datasets, our approach can
 learn transition dynamics and latent states more accurately than competing meth
ods.
****
Optimizing DDPM Sampling with Shortcut Fine-Tuning
Ying Fan, Kangwook Lee
In this study, we propose Shortcut Fine-Tuning (SFT), a new approach for address
ing the challenge of fast sampling of pretrained Denoising Diffusion Probabilist
ic Models (DDPMs). SFT advocates for the fine-tuning of DDPM samplers through th
e direct minimization of Integral Probability Metrics (IPM), instead of learning
 the backward diffusion process. This enables samplers to discover an alternativ
e and more efficient sampling shortcut, deviating from the backward diffusion pr
ocess. Inspired by a control perspective, we propose a new algorithm SFT-PG: Sho
rtcut Fine-Tuning with Policy Gradient, and prove that under certain assumptions
, gradient descent of diffusion models with respect to IPM is equivalent to perf
orming policy gradient. To our best knowledge, this is the first attempt to util
ize reinforcement learning (RL) methods to train diffusion models. Through empir

ical evaluation, we demonstrate that our fine-tuning method can further enhance existing fast DDPM samplers, resulting in sample quality comparable to or even surpassing that of the full-step model across various datasets.
****

LSDS++ : Dual Sampling for Accelerated k-means++
Chenglin Fan, Ping Li, Xiaoyun Li
k-means clustering is an important problem in machine learning and statistics. The k-means++ initialization algorithm has driven new acceleration strategies and theoretical analysis for solving the k-means clustering problem. The state-of-the-art variant, called LocalSearch++, adds extra local search steps upon k-means++ to achieve constant approximation error in expectation. In this paper, we propose a new variant named LSDS++, which improves the sampling efficiency of LocalSearch++ via a strategy called dual sampling. By defining a new capture graph based on the concept of coreset, we show that the proposed LSDS++ is able to achieve the same expected constant error with reduced complexity. Experiments are conducted to justify the benefit of LSDS++ in practice.
****

Smart Initial Basis Selection for Linear Programs
Zhenan Fan, Xinglu Wang, Oleksandr Yakovenko, Abdullah Ali Sivas, Owen Ren, Yong Zhang, Zirui Zhou
The simplex method, introduced by Dantzig more than half a century ago, is still to date one of the most efficient methods for solving large-scale linear programming (LP) problems. While the simplex method is known to have the finite termination property under mild assumptions, the number of iterations until optimality largely depends on the choice of initial basis. Existing strategies for selecting an advanced initial basis are mostly rule-based. These rules usually require extensive expert knowledge and empirical study to develop. Yet, many of them fail to exhibit consistent improvement, even for LP problems that arise in a single application scenario. In this paper, we propose a learning-based approach for initial basis selection. We employ graph neural networks as a building block and develop a model that attempts to capture the relationship between LP problems and their optimal bases. In addition, during the inference phase, we supplement the learning-based prediction with linear algebra tricks to ensure the validity of the generated initial basis. We validate the effectiveness of our proposed strategy by extensively testing it with state-of-the-art simplex solvers, including the open-source solver HiGHS and the commercial solver OptVerse. Through these rigorous experiments, we demonstrate that our strategy achieves substantial speedup and consistently outperforms existing rule-based methods. Furthermore, we extend the proposed approach to generating restricted master problems for column generation methods and present encouraging numerical results.
****

General Covariance Data Augmentation for Neural PDE Solvers
Vladimir Fanaskov, Tianchi Yu, Alexander Rudikov, Ivan Oseledets
The growing body of research shows how to replace classical partial differential equation (PDE) integrators with neural networks. The popular strategy is to generate the input-output pairs with a PDE solver, train the neural network in the regression setting, and use the trained model as a cheap surrogate for the solver. The bottleneck in this scheme is the number of expensive queries of a PDE solver needed to generate the dataset. To alleviate the problem, we propose a computationally cheap augmentation strategy based on general covariance and simple random coordinate transformations. Our approach relies on the fact that physical laws are independent of the coordinate choice, so the change in the coordinate system preserves the type of a parametric PDE and only changes PDE’s data (e.g., initial conditions, diffusion coefficient). For tried neural networks and partial differential equations, proposed augmentation improves test error by 23% on average. The worst observed result is a 17% increase in test error for multilayer perceptron, and the best case is a 80% decrease for dilated residual network.
****

The Fast Johnson-Lindenstrauss Transform Is Even Faster
Ora Nova Fandina, Mikael Møller Høgsgaard, Kasper Green Larsen

The Johnson-Lindenstaruss lemma (Johnson & Lindenstrauss, 1984) is a cornerstone result in dimensionality reduction, stating it is possible to embed a set of n points in d-dimensional Euclidean space into optimal $k=O(\varepsilon^{-2} \ln n)$ dimensions, while preserving all pairwise distances to within a factor $(1 \pm \varepsilon)$. The seminal Fast Johnson-Lindenstrauss (Fast JL) transform by Ailon and Chazelle (SICOMP'09) supports computing the embedding of a data point in $O(d \ln d +k \ln^2 n)$ time, where the $d \ln d$ term comes from multiplication with a $d \times d$ Hadamard matrix and the $k \ln^2 n$ term comes from multiplication with a sparse $k \times d$ matrix. Despite the Fast JL transform being more than a decade old, it is one of the fastest dimensionality reduction techniques for many tradeoffs between ε, d and n. In this work, we give a surprising new analysis of the Fast JL transform, showing that the $k \ln^2 n$ term in the embedding time can be improved to $(k \ln^2 n)/\alpha$ for an $\alpha = \Omega(\min\{\varepsilon^{-1}\ln(1/\varepsilon), \ln n\})$. The improvement follows by using an even sparser matrix. We complement our improved analysis with a lower bound showing that our new analysis is in fact tight.

****

Regression with Label Permutation in Generalized Linear Model

Guanhua Fang, Ping Li

The assumption that response and predictor belong to the same statistical unit may be violated in practice. Unbiased estimation and recovery of true label ordering based on unlabeled data are challenging tasks and have attracted increasing attentions in the recent literature. In this paper, we present a relatively complete analysis of label permutation problem for the generalized linear model with multivariate responses. The theory is established under different scenarios, with knowledge of true parameters, with partial knowledge of underlying label permutation matrix and without any knowledge. Our results remove the stringent conditions required by the current literature and are further extended to the missing observation setting which has never been considered in the field of label permutation problem. On computational side, we propose two methods, "maximum likelihood estimation" algorithm and "two-step estimation" algorithm, to accommodate for different settings. When the proportion of permuted labels is moderate, both methods work effectively. Multiple numerical experiments are provided and corroborate our theoretical findings.

****

Robust Collaborative Learning with Linear Gradient Overhead

Sadegh Farhadkhani, Rachid Guerraoui, Nirupam Gupta, Lê-Nguyên Hoang, Rafael Pinot, John Stephan

Collaborative learning algorithms, such as distributed SGD (or D-SGD), are prone to faulty machines that may deviate from their prescribed algorithm because of software or hardware bugs, poisoned data or malicious behaviors. While many solutions have been proposed to enhance the robustness of D-SGD to such machines, previous works either resort to strong assumptions (trusted server, homogeneous data, specific noise model) or impose a gradient computational cost that is several orders of magnitude higher than that of D-SGD. We present MoNNA, a new algorithm that (a) is provably robust under standard assumptions and (b) has a gradient computation overhead that is linear in the fraction of faulty machines, which is conjectured to be tight. Essentially, MoNNA uses Polyak's momentum of local gradients for local updates and nearest-neighbor averaging (NNA) for global mixing, respectively. While MoNNA is rather simple to implement, its analysis has been more challenging and relies on two key elements that may be of independent interest. Specifically, we introduce the mixing criterion of (α, λ)-reduction to analyze the non-linear mixing of non-faulty machines, and present a way to control the tension between the momentum and the model drifts. We validate our theory by experiments on image classification and make our code available at https://github.com/LPD-EPFL/robust-collaborative-learning.

****

Neural FIM for learning Fisher information metrics from point cloud data

Oluwadamilola Fasina, Guillaume Huguet, Alexander Tong, Yanlei Zhang, Guy Wolf, Maximilian Nickel, Ian Adelstein, Smita Krishnaswamy

Although data diffusion embeddings are ubiquitous in unsupervised learning and h
ave proven to be a viable technique for uncovering the underlying intrinsic geom
etry of data, diffusion embeddings are inherently limited due to their discrete
nature. To this end, we propose neural FIM, a method for computing the Fisher in
formation metric (FIM) from point cloud data - allowing for a continuous manifol
d model for the data. Neural FIM creates an extensible metric space from discret
e point cloud data such that information from the metric can inform us of manifo
ld characteristics such as volume and geodesics. We demonstrate Neural FIM's uti
lity in selecting parameters for the PHATE visualization method as well as its a
bility to obtain information pertaining to local volume illuminating branching p
oints and cluster centers embeddings of a toy dataset and two single-cell datase
ts of IPSC reprogramming and PBMCs (immune cells).
****

Stochastic Policy Gradient Methods: Improved Sample Complexity for Fisher-non-de
generate Policies

Ilyas Fatkhullin, Anas Barakat, Anastasia Kireeva, Niao He

Recently, the impressive empirical success of policy gradient (PG) methods has c
atalyzed the development of their theoretical foundations. Despite the huge effo
rts directed at the design of efficient stochastic PG-type algorithms, the under
standing of their convergence to a globally optimal policy is still limited. In
this work, we develop improved global convergence guarantees for a general class
 of Fisher-non-degenerate parameterized policies which allows to address the cas
e of continuous state action spaces. First, we propose a Normalized Policy Gradi
ent method with Implicit Gradient Transport (N-PG-IGT) and derive a $\tilde{\mat
hcal{O}}(\varepsilon^{-2.5})$ sample complexity of this method for finding a glo
bal ε-optimal policy. Improving over the previously known $\tilde{\m
athcal{O}}(\varepsilon^{-3})$ complexity, this algorithm does not require the us
e of importance sampling or second-order information and samples only one trajec
tory per iteration. Second, we further improve this complexity to $\tilde{ \math
cal{\mathcal{O}} }(\varepsilon^{-2})$ by considering a Hessian-Aided Recursive P
olicy Gradient ((N)-HARPG) algorithm enhanced with a correction based on a Hessi
an-vector product. Interestingly, both algorithms are (i) simple and easy to i
mplement: single-loop, do not require large batches of trajectories and sample a
t most two trajectories per iteration; (ii) computationally and memory efficie
nt: they do not require expensive subroutines at each iteration and can be imple
mented with memory linear in the dimension of parameters.
****

Parallel Neurosymbolic Integration with Concordia

Jonathan Feldstein, Modestas Jur■ius, Efthymia Tsamoura

Parallel neurosymbolic architectures have been applied effectively in NLP by dis
tilling knowledge from a logic theory into a deep model. However, prior art face
s several limitations including supporting restricted forms of logic theories an
d relying on the assumption of independence between the logic and the deep netwo
rk. We present Concordia, a framework overcoming the limitations of prior art. C
oncordia is agnostic both to the deep network and the logic theory offering supp
ort for a wide range of probabilistic theories. Our framework can support superv
ised training of both components and unsupervised training of the neural compone
nt. Concordia has been successfully applied to tasks beyond NLP and data classif
ication, improving the accuracy of state-of-the-art on collective activity detec
tion, entity linking and recommendation tasks.
****

Why Target Networks Stabilise Temporal Difference Methods

Mattie Fellows, Matthew J. A. Smith, Shimon Whiteson

Integral to recent successes in deep reinforcement learning has been a class of
temporal difference methods that use infrequently updated target values for poli
cy evaluation in a Markov Decision Process. Yet a complete theoretical explanati
on for the effectiveness of target networks remains elusive. In this work, we pr
ovide an analysis of this popular class of algorithms, to finally answer the que
stion: "why do target networks stabilise TD learning"? To do so, we formalise th
e notion of a partially fitted policy evaluation method, which describes the use

of target networks and bridges the gap between fitted methods and semigradient temporal difference algorithms. Using this framework we are able to uniquely characterise the so-called deadly triad–the use of TD updates with (nonlinear) function approximation and off-policy data–which often leads to nonconvergent algorithms.This insight leads us to conclude that the use of target networks can mitigate the effects of poor conditioning in the Jacobian of the TD update. Instead, we show that under mild regularity con- ditions and a well tuned target network update frequency, convergence can be guaranteed even in the extremely challenging off-policy sampling and nonlinear function approximation setting.
****

Weighted Sampling without Replacement for Deep Top-k Classification
Dieqiao Feng, Yuanqi Du, Carla P Gomes, Bart Selman
The top-k classification accuracy is a crucial metric in machine learning and is often used to evaluate the performance of deep neural networks. These networks are typically trained using the cross-entropy loss, which optimizes for top-1 classification and is considered optimal in the case of infinite data. However, in real-world scenarios, data is often noisy and limited, leading to the need for more robust losses. In this paper, we propose using the Weighted Sampling Without Replacement (WSWR) method as a learning objective for top-k loss. While traditional methods for evaluating WSWR-based top-k loss are computationally impractical, we show a novel connection between WSWR and Reinforcement Learning (RL) and apply well-established RL algorithms to estimate gradients. We compared our method with recently proposed top-k losses in various regimes of noise and data size for the prevalent use case of $k = 5$. Our experimental results reveal that our method consistently outperforms all other methods on the top-k metric for noisy datasets, has more robustness on extreme testing scenarios, and achieves competitive results on training with limited data.
****

Improved Online Learning Algorithms for CTR Prediction in Ad Auctions
Zhe Feng, Christopher Liaw, Zixin Zhou
In this work, we investigate the online learning problem of revenue maximization in ad auctions, where the seller needs to learn the click-through rates (CTRs) of each ad candidate and charge the price of the winner through a pay-per-click manner. We focus on two models of the advertisers’ strategic behaviors. First, we assume that the advertiser is completely myopic; i.e. in each round, they aim to maximize their utility only for the current round. In this setting, we develop an online mechanism based on upper-confidence bounds that achieves a tight $O(\sqrt{T})$ regret in the worst-case and negative regret when the values are static across all the auctions and there is a gap between the highest expected value (i.e. value multiplied by their CTR) and second highest expected value ad. Next, we assume that the advertiser is non-myopic and cares about their long term utility. This setting is much more complex since an advertiser is incentivized to influence the mechanism by bidding strategically in earlier rounds. In this setting, we provide an algorithm to achieve negative regret for the static valuation setting (with a positive gap), which is in sharp contrast with the prior work that shows $O(T^{2/3})$ regret when the valuation is generated by adversary.
****

Fractional Denoising for 3D Molecular Pre-training
Shikun Feng, Yuyan Ni, Yanyan Lan, Zhi-Ming Ma, Wei-Ying Ma
Coordinate denoising is a promising 3D molecular pre-training method, which has achieved remarkable performance in various downstream drug discovery tasks. Theoretically, the objective is equivalent to learning the force field, which is revealed helpful for downstream tasks. Nevertheless, there are two challenges for coordinate denoising to learn an effective force field, i.e. low coverage samples and isotropic force field. The underlying reason is that molecular distributions assumed by existing denoising methods fail to capture the anisotropic characteristic of molecules. To tackle these challenges, we propose a novel hybrid noise strategy, including noises on both dihedral angel and coordinate. However, denoising such hybrid noise in a traditional way is no more equivalent to learning the force field. Through theoretical deductions, we find that the problem is caus

ed by the dependency of the input conformation for covariance. To this end, we p
ropose to decouple the two types of noise and design a novel fractional denoisin
g method (Frad), which only denoises the latter coordinate part. In this way, Fr
ad enjoys both the merits of sampling more low-energy structures and the force f
ield equivalence. Extensive experiments show the effectiveness of Frad in molecu
le representation, with a new state-of-the-art on 9 out of 12 tasks of QM9 and o
n 7 out of 8 targets of MD17.
****

Improved Algorithms for White-Box Adversarial Streams
Ying Feng, David Woodruff
We study streaming algorithms in the white-box adversarial stream model, where t
he internal state of the streaming algorithm is revealed to an adversary who ada
ptively generates the stream updates, but the algorithm obtains fresh randomness
 unknown to the adversary at each time step. We incorporate cryptographic assump
tions to construct robust algorithms against such adversaries. We propose effici
ent algorithms for sparse recovery of vectors, low rank recovery of matrices and
 tensors, as well as low rank plus sparse recovery of matrices, i.e., robust PCA
. Unlike deterministic algorithms, our algorithms can report when the input is n
ot sparse or low rank even in the presence of such an adversary. We use these re
covery algorithms to improve upon and solve new problems in numerical linear alg
ebra and combinatorial optimization on white-box adversarial streams. For exampl
e, we give the first efficient algorithm for outputting a matching in a graph wi
th insertions and deletions to its edges provided the matching size is small, an
d otherwise we declare the matching size is large. We also improve the approxima
tion versus memory tradeoff of previous work for estimating the number of non-ze
ro elements in a vector and computing the matrix rank.
****

Non-stationary Reinforcement Learning under General Function Approximation
Songtao Feng, Ming Yin, Ruiquan Huang, Yu-Xiang Wang, Jing Yang, Yingbin Liang
General function approximation is a powerful tool to handle large state and acti
on spaces in a broad range of reinforcement learning (RL) scenarios. However, th
eoretical understanding of non-stationary MDPs with general function approximati
on is still limited. In this paper, we make the first such an attempt. We first
propose a new complexity metric called dynamic Bellman Eluder (DBE) dimension fo
r non-stationary MDPs, which subsumes majority of existing tractable RL problems
 in static MDPs as well as non-stationary MDPs. Based on the proposed complexity
 metric, we propose a novel confidence-set based model-free algorithm called SW-
OPEA, which features a sliding window mechanism and a new confidence set design
for non-stationary MDPs. We then establish an upper bound on the dynamic regret
for the proposed algorithm, and show that SW-OPEA is provably efficient as long
as the variation budget is not significantly large. We further demonstrate via e
xamples of non-stationary linear and tabular MDPs that our algorithm performs be
tter in small variation budget scenario than the existing UCB-type algorithms. T
o the best of our knowledge, this is the first dynamic regret analysis in non-st
ationary MDPs with general function approximation.
****

Random Matrix Analysis to Balance between Supervised and Unsupervised Learning u
nder the Low Density Separation Assumption
Vasilii Feofanov, Malik Tiomoko, Aladin Virmaux
We propose a theoretical framework to analyze semi-supervised classification und
er the low density separation assumption in a high-dimensional regime. In partic
ular, we introduce QLDS, a linear classification model, where the low density se
paration assumption is implemented via quadratic margin maximization. The algori
thm has an explicit solution with rich theoretical properties, and we show that
particular cases of our algorithm are the least-square support vector machine in
 the supervised case, the spectral clustering in the fully unsupervised regime,
and a class of semi-supervised graph-based approaches. As such, QLDS establishes
 a smooth bridge between these supervised and unsupervised learning methods. Usi
ng recent advances in the random matrix theory, we formally derive a theoretical
 evaluation of the classification error in the asymptotic regime. As an applicat

ion, we derive a hyperparameter selection policy that finds the best balance between the supervised and the unsupervised terms of our learning criterion. Finally, we provide extensive illustrations of our framework, as well as an experimental study on several benchmarks to demonstrate that QLDS, while being computationally more efficient, improves over cross-validation for hyperparameter selection, indicating a high promise of the usage of random matrix theory for semi-supervised model selection.
****
SurCo: Learning Linear SURrogates for COmbinatorial Nonlinear Optimization Problems
Aaron M Ferber, Taoan Huang, Daochen Zha, Martin Schubert, Benoit Steiner, Bistra Dilkina, Yuandong Tian
Optimization problems with nonlinear cost functions and combinatorial constraints appear in many real-world applications but remain challenging to solve efficiently compared to their linear counterparts. To bridge this gap, we propose $\textbf{\emph{\texttt{SurCo}}}$ that learns linear $\underline{\text{Sur}}$rogate costs which can be used in existing $\underline{\text{Co}}$mbinatorial solvers to output good solutions to the original nonlinear combinatorial optimization problem. The surrogate costs are learned end-to-end with nonlinear loss by differentiating through the linear surrogate solver, combining the flexibility of gradient-based methods with the structure of linear combinatorial optimization. We propose three $\texttt{SurCo}$ variants: $\texttt{SurCo}-\texttt{zero}$ for individual nonlinear problems, $\texttt{SurCo}-\texttt{prior}$ for problem distributions, and $\texttt{SurCo}-\texttt{hybrid}$ to combine both distribution and problem-specific information. We give theoretical intuition motivating $\texttt{SurCo}$, and evaluate it empirically. Experiments show that $\texttt{SurCo}$ finds better solutions faster than state-of-the-art and domain expert approaches in real-world optimization problems such as embedding table sharding, inverse photonic design, and nonlinear route planning.
****
Scaling Laws for Multilingual Neural Machine Translation
Patrick Fernandes, Behrooz Ghorbani, Xavier Garcia, Markus Freitag, Orhan Firat
In this work, we provide a large-scale empirical study of the scaling properties of multilingual neural machine translation models. We examine how increases in the model size affect the model performance and investigate the role of the individual language pair weights on the scaling behavior. We find that these weights only affect the multiplicative factor of the scaling law, and in particular, the scaling exponent is unaffected by them. Through a novel joint scaling law formulation, we compute the effective number of parameters allocated to each language pair and examine the role of language similarity in the scaling behavior of our models. We find little evidence that language similarity has any impact. In contrast, “direction” of the multilinguality plays a significant role, with models translating from multiple languages into English having a larger number of effective parameters per task than their reversed counterparts. Finally, we leverage our observations to predict the performance of multilingual models trained with any language weighting at any scale, greatly reducing efforts required for language balancing in large multilingual models. Our findings apply to both in-domain and out-of-domain test sets and to multiple evaluation metrics, such as ChrF and BLEURT.
****
Constant Matters: Fine-grained Error Bound on Differentially Private Continual Observation
Hendrik Fichtenberger, Monika Henzinger, Jalaj Upadhyay
We study fine-grained error bounds for differentially private algorithms for counting under continual observation. Our main insight is that the matrix mechanism when using lower-triangular matrices can be used in the continual observation model. More specifically, we give an explicit factorization for the counting matrix M_count and upper bound the error explicitly. We also give a fine-grained analysis, specifying the exact constant in the upper bound. Our analysis is based on upper and lower bounds of the completely bounded norm (cb-norm) of

M_count. Along the way, we improve the best-known bound of 28 years by Mathias (SIAM Journal on Matrix Analysis and Applications, 1993) on the cb-norm of M_count for a large range of the dimension of M_count. Furthermore, we are the first to give concrete error bounds for various problems under continual observation such as binary counting, maintaining a histogram, releasing an approximately cut-preserving synthetic graph, many graph-based statistics, and substring and episode counting. Finally, we note that our result can be used to get a fine-grained error bound for non-interactive local learning and the first lower bounds on the additive error for (ϵ,δ)-differentially-private counting under continual observation. Subsequent to this work, Henzinger et al. (SODA, 2023) showed that our factorization also achieves fine-grained mean-squared error.

****

Adapting to game trees in zero-sum imperfect information games

Côme Fiegel, Pierre Menard, Tadashi Kozuno, Remi Munos, Vianney Perchet, Michal Valko

Imperfect information games (IIG) are games in which each player only partially observes the current game state. We study how to learn ϵ-optimal strategies in a zero-sum IIG through self-play with trajectory feedback. We give a problem-independent lower bound $\widetilde{\mathcal{O}}(H(A_{\mathcal{X}}+B_{\mathcal{Y}})/\epsilon^2)$ on the required number of realizations to learn these strategies with high probability, where H is the length of the game, $A_{\mathcal{X}}$ and $B_{\mathcal{Y}}$ are the total number of actions for the two players. We also propose two Follow the Regularized leader (FTRL) algorithms for this setting: Balanced FTRL which matches this lower bound, but requires the knowledge of the information set structure beforehand to define the regularization; and Adaptive FTRL which needs $\widetilde{\mathcal{O}}(H^2(A_{\mathcal{X}}+B_{\mathcal{Y}})/\epsilon^2)$ realizations without this requirement by progressively adapting the regularization to the observations.

****

User-defined Event Sampling and Uncertainty Quantification in Diffusion Models for Physical Dynamical Systems

Marc Anton Finzi, Anudhyan Boral, Andrew Gordon Wilson, Fei Sha, Leonardo Zepeda-Nunez

Diffusion models are a class of probabilistic generative models that have been widely used as a prior for image processing tasks like text conditional generation and inpainting. We demonstrate that these models can be adapted to make predictions and provide uncertainty quantification for chaotic dynamical systems. In these applications, diffusion models can implicitly represent knowledge about outliers and extreme events; however, querying that knowledge through conditional sampling or measuring probabilities is surprisingly difficult. Existing methods for conditional sampling at inference time seek mainly to enforce the constraints, which is insufficient to match the statistics of the distribution or compute the probability of the chosen events. To achieve these ends, optimally one would use the conditional score function, but its computation is typically intractable. In this work, we develop a probabilistic approximation scheme for the conditional score function which provably converges to the true distribution as the noise level decreases. With this scheme we are able to sample conditionally on nonlinear user-defined events at inference time, and matches data statistics even when sampling from the tails of the distribution.

****

ACAT: Adversarial Counterfactual Attention for Classification and Detection in Medical Imaging

Alessandro Fontanella, Antreas Antoniou, Wenwen Li, Joanna Wardlaw, Grant Mair, Emanuele Trucco, Amos Storkey

In some medical imaging tasks and other settings where only small parts of the image are informative for the classification task, traditional CNNs can sometimes struggle to generalise. Manually annotated Regions of Interest (ROI) are often used to isolate the most informative parts of the image. However, these are expensive to collect and may vary significantly across annotators. To overcome these

issues, we propose a framework that employs saliency maps to obtain soft spatial attention masks that modulate the image features at different scales. We refer to our method as Adversarial Counterfactual Attention (ACAT). ACAT increases the baseline classification accuracy of lesions in brain CT scans from $71.39 %$ to $72.55 %$ and of COVID-19 related findings in lung CT scans from $67.71 %$ to $70.84 %$ and exceeds the performance of competing methods. We investigate the best way to generate the saliency maps employed in our architecture and propose a way to obtain them from adversarially generated counterfactual images. They are able to isolate the area of interest in brain and lung CT scans without using any manual annotations. In the task of localising the lesion location out of 6 possible regions, they obtain a score of $65.05 %$ on brain CT scans, improving the score of $61.29 %$ obtained with the best competing method.

****

Explainable Data-Driven Optimization: From Context to Decision and Back Again

Alexandre Forel, Axel Parmentier, Thibaut Vidal

Data-driven optimization uses contextual information and machine learning algorithms to find solutions to decision problems with uncertain parameters. While a vast body of work is dedicated to interpreting machine learning models in the classification setting, explaining decision pipelines involving learning algorithms remains unaddressed. This lack of interpretability can block the adoption of data-driven solutions as practitioners may not understand or trust the recommended decisions. We bridge this gap by introducing a counterfactual explanation methodology tailored to explain solutions to data-driven problems. We introduce two classes of explanations and develop methods to find nearest explanations of random forest and nearest-neighbor predictors. We demonstrate our approach by explaining key problems in operations management such as inventory management and routing.

****

Hardness of Independent Learning and Sparse Equilibrium Computation in Markov Games

Dylan J Foster, Noah Golowich, Sham M. Kakade

We consider the problem of decentralized multi-agent reinforcement learning in Markov games. A fundamental question is whether there exist algorithms that, when run independently by all agents, lead to no-regret for each player, analogous to celebrated convergence results for no-regret learning in normal-form games. While recent work has shown that such algorithms exist for restricted settings (notably, when regret is defined with respect to deviations to Markov policies), the question of whether independent no-regret learning can be achieved in the standard Markov game framework was open. We provide a decisive negative resolution to this problem, both from a computational and statistical perspective. We show that: • Under the complexity-theoretic assumption that PPAD $\neq$ P, there is no polynomial-time algorithm that attains no-regret in two-player general-sum Markov games when executed independently by all players, even when the game is known to the algorithm designer. • When the game is unknown, no algorithm, efficient or otherwise, can achieve no-regret without observing exponentially many episodes in the number of players. These results are proven via lower bounds for a simpler problem we refer to as SparseCCE, in which the goal is to compute a coarse correlated equilibrium that is “sparse” in the sense that it can be represented as a mixture of a small number of product policies.

****

Disentangled Generative Models for Robust Prediction of System Dynamics

Stathi Fotiadis, Mario Lino Valencia, Shunlong Hu, Stef Garasto, Chris D Cantwell, Anil Anthony Bharath

The use of deep neural networks for modelling system dynamics is increasingly popular, but long-term prediction accuracy and out-of-distribution generalization still present challenges. In this study, we address these challenges by considering the parameters of dynamical systems as factors of variation of the data and leverage their ground-truth values to disentangle the representations learned by generative models. Our experimental results in phase-space and observation-space dynamics, demonstrate the effectiveness of latent-space supervision in produci

ng disentangled representations, leading to improved long-term prediction accura
cy and out-of-distribution robustness.
****
Can Forward Gradient Match Backpropagation?
Louis Fournier, Stephane Rivaud, Eugene Belilovsky, Michael Eickenberg, Edouard
Oyallon
Forward Gradients - the idea of using directional derivatives in forward differe
ntiation mode - have recently been shown to be utilizable for neural network tra
ining while avoiding problems generally associated with backpropagation gradient
 computation, such as locking and memorization requirements. The cost is the req
uirement to guess the step direction, which is hard in high dimensions. While cu
rrent solutions rely on weighted averages over isotropic guess vector distributi
ons, we propose to strongly bias our gradient guesses in directions that are muc
h more promising, such as feedback obtained from small, local auxiliary networks
. For a standard computer vision neural network, we conduct a rigorous study sys
tematically covering a variety of combinations of gradient targets and gradient
guesses, including those previously presented in the literature. We find that us
ing gradients obtained from a local loss as a candidate direction drastically im
proves on random noise in Forward Gradient methods.
****
Last Switch Dependent Bandits with Monotone Payoff Functions
Ayoub Foussoul, Vineet Goyal, Orestis Papadigenopoulos, Assaf Zeevi
In a recent work, Laforgue et al. introduce the model of last switch dependent (
LSD) bandits, in an attempt to capture nonstationary phenomena induced by the in
teraction between the player and the environment. Examples include satiation, wh
ere consecutive plays of the same action lead to decreased performance, or depri
vation, where the payoff of an action increases after an interval of inactivity.
 In this work, we take a step towards understanding the approximability of plann
ing LSD bandits, namely, the (NP-hard) problem of computing an optimal arm-pulli
ng strategy under complete knowledge of the model. In particular, we design the
first efficient constant approximation algorithm for the problem and show that,
under a natural monotonicity assumption on the payoffs, its approximation guaran
tee (almost) matches the state-of-the-art for the special and well-studied class
 of recharging bandits (also known as delay-dependent). In this attempt, we deve
lop new tools and insights for this class of problems, including a novel higher-
dimensional relaxation and the technique of mirroring the evolution of virtual s
tates. We believe that these novel elements could potentially be used for approa
ching richer classes of action-induced nonstationary bandits (e.g., special inst
ances of restless bandits). In the case where the model parameters are initially
 unknown, we develop an online learning adaptation of our algorithm for which we
 provide sublinear regret guarantees against its full-information counterpart.
****
A Theoretical Analysis of the Learning Dynamics under Class Imbalance
Emanuele Francazi, Marco Baity-Jesi, Aurelien Lucchi
Data imbalance is a common problem in machine learning that can have a critical
effect on the performance of a model. Various solutions exist but their impact o
n the convergence of the learning dynamics is not understood. Here, we elucidate
 the significant negative impact of data imbalance on learning, showing that the
 learning curves for minority and majority classes follow sub-optimal trajectori
es when training with a gradient-based optimizer. This slowdown is related to th
e imbalance ratio and can be traced back to a competition between the optimizati
on of different classes. Our main contribution is the analysis of the convergenc
e of full-batch (GD) and stochastic gradient descent (SGD), and of variants that
 renormalize the contribution of each per-class gradient. We find that GD is not
 guaranteed to decrease the loss for each class but that this problem can be add
ressed by performing a per-class normalization of the gradient. With SGD, class
imbalance has an additional effect on the direction of the gradients: the minori
ty class suffers from a higher directional noise, which reduces the effectivenes
s of the per-class gradient normalization. Our findings not only allow us to und
erstand the potential and limitations of strategies involving the per-class grad

ients, but also the reason for the effectiveness of previously used solutions fo
r class imbalancesuch as oversampling.
****

SparseGPT: Massive Language Models Can be Accurately Pruned in One-Shot
Elias Frantar, Dan Alistarh
We show for the first time that large-scale generative pretrained transformer (G
PT) family models can be pruned to at least 50% sparsity in one-shot, without an
y retraining, at minimal loss of accuracy. This is achieved via a new pruning me
thod called SparseGPT, specifically designed to work efficiently and accurately
on massive GPT-family models. We can execute SparseGPT on the largest available
open-source models, OPT-175B and BLOOM-176B, in under 4.5 hours, and can reach 6
0% unstructured sparsity with negligible increase in perplexity: remarkably, mor
e than 100 billion weights from these models can be ignored at inference time. S
parseGPT generalizes to semi-structured (2:4 and 4:8) patterns, and is compatibl
e with weight quantization approaches. The code is available at: https://github.
com/IST-DASLab/sparsegpt.
****

Learning Temporally AbstractWorld Models without Online Experimentation
Benjamin Freed, Siddarth Venkatraman, Guillaume Adrien Sartoretti, Jeff Schneide
r, Howie Choset
Agents that can build temporally abstract representations of their environment a
re better able to understand their world and make plans on extended time scales,
 with limited computational power and modeling capacity. However, existing metho
ds for automatically learning temporally abstract world models usually require m
illions of online environmental interactions and incentivize agents to reach eve
ry accessible environmental state, which is infeasible for most real-world robot
s both in terms of data efficiency and hardware safety. In this paper, we presen
t an approach for simultaneously learning sets of skills and temporally abstract
, skill-conditioned world models purely from offline data, enabling agents to pe
rform zero-shot online planning of skill sequences for new tasks. We show that o
ur approach performs comparably to or better than a wide array of state-of-the-a
rt offline RL algorithms on a number of simulated robotics locomotion and manipu
lation benchmarks, while offering a higher degree of adaptability to new goals.
Finally, we show that our approach offers a much higher degree of robustness to
perturbations in environmental dynamics, compared to policy-based methods.
****

A Coupled Flow Approach to Imitation Learning
Gideon Joseph Freund, Elad Sarafian, Sarit Kraus
In reinforcement learning and imitation learning, an object of central importanc
e is the state distribution induced by the policy. It plays a crucial role in th
e policy gradient theorem, and references to it–along with the related state-act
ion distribution–can be found all across the literature. Despite its importance,
 the state distribution is mostly discussed indirectly and theoretically, rather
 than being modeled explicitly. The reason being an absence of appropriate densi
ty estimation tools. In this work, we investigate applications of a normalizing
flow based model for the aforementioned distributions. In particular, we use a p
air of flows coupled through the optimality point of the Donsker-Varadhan repres
entation of the Kullback-Leibler (KL) divergence, for distribution matching base
d imitation learning. Our algorithm, Coupled Flow Imitation Learning (CFIL), ach
ieves state-of-the-art performance on benchmark tasks with a single expert traje
ctory and extends naturally to a variety of other settings, including the subsam
pled and state-only regimes.
****

Simple Hardware-Efficient Long Convolutions for Sequence Modeling
Daniel Y Fu, Elliot L Epstein, Eric Nguyen, Armin W Thomas, Michael Zhang, Tri D
ao, Atri Rudra, Christopher Re
State space models (SSMs) have high performance on long sequence modeling but re
quire sophisticated initialization techniques and specialized implementations fo
r high quality and runtime performance. We study whether a simple alternative ca
n match SSMs in performance and efficiency: directly learning long convolutions

over the sequence. We find that a key requirement to achieving high performance is keeping the convolution kernels smooth. We find that simple interventions-such as squashing the kernel weights-result in smooth kernels and recover SSM performance on a range of tasks including the long range arena, image classification, language modeling, and brain data modeling. Next, we develop FlashButterfly, an IO-aware algorithm to improve the runtime performance of long convolutions. FlashButterfly appeals to classic Butterfly decompositions of the convolution to reduce GPU memory IO and increase FLOP utilization. FlashButterfly speeds up convolutions by 2.2$\times$, and allows us to train on Path256, a challenging task with sequence length 64K, where we set state-of-the-art by 29.1 points while training 7.2$\times$ faster than prior work. Lastly, we introduce an extension to FlashButterfly that learns the coefficients of the Butterfly decomposition, increasing expressivity without increasing runtime. Using this extension, we outperform a Transformer on WikiText103 by 0.2 PPL with 30% fewer parameters.
****
MonoNeRF: Learning Generalizable NeRFs from Monocular Videos without Camera Poses
Yang Fu, Ishan Misra, Xiaolong Wang
We propose a generalizable neural radiance fields - MonoNeRF, that can be trained on large-scale monocular videos of moving in static scenes without any ground-truth annotations of depth and camera poses. MonoNeRF follows an Autoencoder-based architecture, where the encoder estimates the monocular depth and the camera pose, and the decoder constructs a Multiplane NeRF representation based on the depth encoder feature, and renders the input frames with the estimated camera. The learning is supervised by the reconstruction error. Once the model is learned, it can be applied to multiple applications including depth estimation, camera pose estimation, and single-image novel view synthesis. More qualitative results are available at: https://oasisyang.github.io/mononerf.
****
Go Beyond Imagination: Maximizing Episodic Reachability with World Models
Yao Fu, Run Peng, Honglak Lee
Efficient exploration is a challenging topic in reinforcement learning, especially for sparse reward tasks. To deal with the reward sparsity, people commonly apply intrinsic rewards to motivate agents to explore the state space efficiently. In this paper, we introduce a new intrinsic reward design called GoBI - Go Beyond Imagination, which combines the traditional lifelong novelty motivation with an episodic intrinsic reward that is designed to maximize the stepwise reachability expansion. More specifically, we apply learned world models to generate predicted future states with random actions. States with more unique predictions that are not in episodic memory are assigned high intrinsic rewards. Our method greatly outperforms previous state-of-the-art methods on 12 of the most challenging Minigrid navigation tasks and improves the sample efficiency on locomotion tasks from DeepMind Control Suite.
****
Specializing Smaller Language Models towards Multi-Step Reasoning
Yao Fu, Hao Peng, Litu Ou, Ashish Sabharwal, Tushar Khot
The surprising ability of Large Language Models (LLMs) to perform well on complex reasoning with only few-shot chain-of-thought prompts is believed to emerge only in very large-scale models. We show that such abilities can, in fact, be distilled down from GPT-3.5 (≥ 175B) to T5 variants (≤ 11B). We propose model specialization, to specialize the model’s ability towards a target task. The hypothesis is that large models (commonly viewed as larger than 100B) have strong modeling power such that they can perform a large spectrum of tasks. Small models (commonly viewed as smaller than 10B) have limited model capacity, but if we specialize their capacity towards a target task, the model can achieve decent performance improvements. We use multi-step math reasoning as our testbed because it is a very typical emergent ability. We show two important aspects of model abilities: (1) balancing language model’s performance on multiple tasks is a delicate matter, as improvements on one task may compromise other tasks; (2) yet by intentionally paying the price of decreased generic ability, we can clearly improve acros

s different model scales smaller than 10B towards a specialized multi-step math reasoning ability. We further give comprehensive discussions about important design choices for better generalization, including the data format mixture and the start model checkpoint. We hope our practice and discoveries can serve as an important attempt towards specialized smaller models in the new research paradigm set by LLMs.
****

Accelerated Stochastic Optimization Methods under Quasar-convexity
Qiang Fu, Dongchu Xu, Ashia Camage Wilson
Non-convex optimization plays a key role in a growing number of machine learning applications. This motivates the identification of specialized structure that enables sharper theoretical analysis. One such identified structure is quasar-convexity, a non-convex generalization of convexity that subsumes convex functions. Existing algorithms for minimizing quasar-convex functions in the stochastic setting have either high complexity or slow convergence, which prompts us to derive a new class of stochastic methods for optimizing smooth quasar-convex functions. We demonstrate that our algorithms have fast convergence and outperform existing algorithms on several examples, including the classical problem of learning linear dynamical systems. We also present a unified analysis of our newly proposed algorithms and a previously studied deterministic algorithm.
****

Meta-learning Parameterized Skills
Haotian Fu, Shangqun Yu, Saket Tiwari, Michael Littman, George Konidaris
We propose a novel parameterized skill-learning algorithm that aims to learn transferable parameterized skills and synthesize them into a new action space that supports efficient learning in long-horizon tasks. We propose to leverage off-policy Meta-RL combined with a trajectory-centric smoothness term to learn a set of parameterized skills. Our agent can use these learned skills to construct a three-level hierarchical framework that models a Temporally-extended Parameterized Action Markov Decision Process. We empirically demonstrate that the proposed algorithms enable an agent to solve a set of highly difficult long-horizon (obstacle-course and robot manipulation) tasks.
****

NeRFool: Uncovering the Vulnerability of Generalizable Neural Radiance Fields against Adversarial Perturbations
Yonggan Fu, Ye Yuan, Souvik Kundu, Shang Wu, Shunyao Zhang, Yingyan Celine Lin
Generalizable Neural Radiance Fields (GNeRF) are one of the most promising real-world solutions for novel view synthesis, thanks to their cross-scene generalization capability and thus the possibility of instant rendering on new scenes. While adversarial robustness is essential for real-world applications, little study has been devoted to understanding its implication on GNeRF. We hypothesize that because GNeRF is implemented by conditioning on the source views from new scenes, which are often acquired from the Internet or third-party providers, there are potential new security concerns regarding its real-world applications. Meanwhile, existing understanding and solutions for neural networks' adversarial robustness may not be applicable to GNeRF, due to its 3D nature and uniquely diverse operations. To this end, we present NeRFool, which to the best of our knowledge is the first work that sets out to understand the adversarial robustness of GNeRF. Specifically, NeRFool unveils the vulnerability patterns and important insights regarding GNeRF's adversarial robustness. Built upon the above insights gained from NeRFool, we further develop NeRFool$^+$, which integrates two techniques capable of effectively attacking GNeRF across a wide range of target views, and provide guidelines for defending against our proposed attacks. We believe that our NeRFool/NeRFool$^+$ lays the initial foundation for future innovations in developing robust real-world GNeRF solutions. Our codes are available at: https://github.com/GATECH-EIC/NeRFool.
****

Hierarchies of Reward Machines
Daniel Furelos-Blanco, Mark Law, Anders Jonsson, Krysia Broda, Alessandra Russo
Reward machines (RMs) are a recent formalism for representing the reward functio

n of a reinforcement learning task through a finite-state machine whose edges en
code subgoals of the task using high-level events. The structure of RMs enables
the decomposition of a task into simpler and independently solvable subtasks tha
t help tackle long-horizon and/or sparse reward tasks. We propose a formalism fo
r further abstracting the subtask structure by endowing an RM with the ability t
o call other RMs, thus composing a hierarchy of RMs (HRM). We exploit HRMs by tr
eating each call to an RM as an independently solvable subtask using the options
 framework, and describe a curriculum-based method to learn HRMs from traces obs
erved by the agent. Our experiments reveal that exploiting a handcrafted HRM lea
ds to faster convergence than with a flat HRM, and that learning an HRM is feasi
ble in cases where its equivalent flat representation is not.
****

Why Random Pruning Is All We Need to Start Sparse
Advait Harshal Gadhikar, Sohom Mukherjee, Rebekka Burkholz
Random masks define surprisingly effective sparse neural network models, as has
been shown empirically. The resulting sparse networks can often compete with den
se architectures and state-of-the-art lottery ticket pruning algorithms, even th
ough they do not rely on computationally expensive prune-train iterations and ca
n be drawn initially without significant computational overhead. We offer a theo
retical explanation of how random masks can approximate arbitrary target network
s if they are wider by a logarithmic factor in the inverse sparsity $1 / \log(1/
\text{sparsity})$. This overparameterization factor is necessary at least for 3-
layer random networks, which elucidates the observed degrading performance of ra
ndom networks at higher sparsity. At moderate to high sparsity levels, however,
our results imply that sparser networks are contained within random source netwo
rks so that any dense-to-sparse training scheme can be turned into a computation
ally more efficient sparse-to-sparse one by constraining the search to a fixed r
andom mask. We demonstrate the feasibility of this approach in experiments for d
ifferent pruning methods and propose particularly effective choices of initial l
ayer-wise sparsity ratios of the random source network. As a special case, we sh
ow theoretically and experimentally that random source networks also contain str
ong lottery tickets.
****

Cell-Free Latent Go-Explore
Quentin Gallouédec, Emmanuel Dellandrea
In this paper, we introduce Latent Go-Explore (LGE), a simple and general approa
ch based on the Go-Explore paradigm for exploration in reinforcement learning (R
L). Go-Explore was initially introduced with a strong domain knowledge constrain
t for partitioning the state space into cells. However, in most real-world scena
rios, drawing domain knowledge from raw observations is complex and tedious. If
the cell partitioning is not informative enough, Go-Explore can completely fail
to explore the environment. We argue that the Go-Explore approach can be general
ized to any environment without domain knowledge and without cells by exploiting
 a learned latent representation. Thus, we show that LGE can be flexibly combine
d with any strategy for learning a latent representation. Our results indicate t
hat LGE, although simpler than Go-Explore, is more robust and outperforms state-
of-the-art algorithms in terms of pure exploration on multiple hard-exploration
environments including Montezuma’s Revenge. The LGE implementation is available
as open-source at https://github.com/qgallouedec/lge.
****

Graph Reinforcement Learning for Network Control via Bi-Level Optimization
Daniele Gammelli, James Harrison, Kaidi Yang, Marco Pavone, Filipe Rodrigues, Fr
ancisco C. Pereira
Optimization problems over dynamic networks have been extensively studied and wi
dely used in the past decades to formulate numerous real-world problems. However
, (1) traditional optimization-based approaches do not scale to large networks,
and (2) the design of good heuristics or approximation algorithms often requires
 significant manual trial-and-error. In this work, we argue that data-driven str
ategies can automate this process and learn efficient algorithms without comprom
ising optimality. To do so, we present network control problems through the lens

of reinforcement learning and propose a graph network-based framework to handle a broad class of problems. Instead of naively computing actions over high-dimensional graph elements, e.g., edges, we propose a bi-level formulation where we (1) specify a desired next state via RL, and (2) solve a convex program to best achieve it, leading to drastically improved scalability and performance. We further highlight a collection of desirable features to system designers, investigate design decisions, and present experiments on real-world control problems showing the utility, scalability, and flexibility of our framework.
****
Why Is Public Pretraining Necessary for Private Model Training?
Arun Ganesh, Mahdi Haghifam, Milad Nasr, Sewoong Oh, Thomas Steinke, Om Thakkar, Abhradeep Guha Thakurta, Lun Wang
In the privacy-utility tradeoff of a model trained on benchmark language and vision tasks, remarkable improvements have been widely reported when the model is pretrained on public data. Some gain is expected as these models inherit the benefits of transfer learning, which is the standard motivation in non-private settings. However, the stark contrast in the gain of pretraining between non-private and private machine learning suggests that the gain in the latter is rooted in a fundamentally different cause. To explain this phenomenon, we hypothesize that the non-convex loss landscape of a model training necessitates the optimization algorithm to go through two phases. In the first, the algorithm needs to select a good “basin” in the loss landscape. In the second, the algorithm solves an easy optimization within that basin. The former is a harder problem to solve with private data, while the latter is harder to solve with public data due to a distribution shift or data scarcity. Guided by this intuition, we provide theoretical constructions that provably demonstrate the separation between private training with and without public pretraining. Further, systematic experiments on CIFAR10 and Librispeech provide supporting evidence for our hypothesis.
****
Do Perceptually Aligned Gradients Imply Robustness?
Roy Ganz, Bahjat Kawar, Michael Elad
Adversarially robust classifiers possess a trait that non-robust models do not - Perceptually Aligned Gradients (PAG). Their gradients with respect to the input align well with human perception. Several works have identified PAG as a byproduct of robust training, but none have considered it as a standalone phenomenon nor studied its own implications. In this work, we focus on this trait and test whether Perceptually Aligned Gradients imply Robustness. To this end, we develop a novel objective to directly promote PAG in training classifiers and examine whether models with such gradients are more robust to adversarial attacks. Extensive experiments on multiple datasets and architectures validate that models with aligned gradients exhibit significant robustness, exposing the surprising bidirectional connection between PAG and robustness. Lastly, we show that better gradient alignment leads to increased robustness and harness this observation to boost the robustness of existing adversarial training techniques.
****
Solving Linear Programs with Fast Online Learning Algorithms
Wenzhi Gao, Dongdong Ge, Chunlin Sun, Yinyu Ye
This paper presents fast first-order methods for solving linear programs (LPs) approximately. We adapt online linear programming algorithms to offline LPs and obtain algorithms that avoid any matrix multiplication. We also introduce a variable-duplication technique that copies each variable K times and reduces the optimality gap and constraint violation by a factor of $\sqrt{K}$. Furthermore, we show how online algorithms can be effectively integrated into sifting, a column generation scheme for large-scale LPs. Numerical experiments demonstrate that our methods can serve as either an approximate direct solver, or an initialization subroutine for exact LP solving.
****
Gradient Descent Finds the Global Optima of Two-Layer Physics-Informed Neural Networks
Yihang Gao, Yiqi Gu, Michael Ng

The main aim of this paper is to conduct the convergence analysis of the gradien t descent for two-layer physics-informed neural networks (PINNs). Here, the loss function involves derivatives of neural network outputs with respect to its inp uts, so the interaction between the trainable parameters is more complicated com pared with simple regression and classification tasks. We first develop the posi tive definiteness of Gram matrices and prove that the gradient flow finds the gl obal optima of the empirical loss under over-parameterization. Then, we demonstr ate that the standard gradient descent converges to the global optima of the los s with proper choices of learning rates. The framework of our analysis works for various categories of PDEs (e.g., linear second-order PDEs) and common types of network initialization (LecunUniform etc.). Our theoretical results do not need a very strict hypothesis for training samples and have a looser requirement on the network width compared with some previous works.

****

Generalizing Neural Wave Functions

Nicholas Gao, Stephan Günnemann

Recent neural network-based wave functions have achieved state-of-the-art accura cies in modeling ab-initio ground-state potential energy surface. However, these networks can only solve different spatial arrangements of the same set of atoms . To overcome this limitation, we present Graph-learned orbital embeddings (Glob e), a neural network-based reparametrization method that can adapt neural wave f unctions to different molecules. Globe learns representations of local electroni c structures that generalize across molecules via spatial message passing by con necting molecular orbitals to covalent bonds. Further, we propose a size-consist ent wave function Ansatz, the Molecular orbital network (Moon), tailored to join tly solve Schrödinger equations of different molecules. In our experiments, we f ind Moon converging in 4.5 times fewer steps to similar accuracy as previous met hods or to lower energies given the same time. Further, our analysis shows that Moon’s energy estimate scales additively with increased system sizes, unlike pre vious work where we observe divergence. In both computational chemistry and mach ine learning, we are the first to demonstrate that a single wave function can so lve the Schrödinger equation of molecules with different atoms jointly.

****

On the Impact of Algorithmic Recourse on Social Segregation

Ruijiang Gao, Himabindu Lakkaraju

As predictive models seep into several real-world applications, it has become cr itical to ensure that individuals who are negatively impacted by the outcomes of these models are provided with a means for recourse. To this end, there has bee n a growing body of research on algorithmic recourse in recent years. While reco urses can be extremely beneficial to affected individuals, their implementation at a large scale can lead to potential data distribution shifts and other uninte nded consequences. However, there is little to no research on understanding the impact of algorithmic recourse after implementation. In this work, we address th e aforementioned gaps by making one of the first attempts at analyzing the delay ed societal impact of algorithmic recourse. To this end, we theoretically and em pirically analyze the recourses output by state-of-the-art algorithms. Our analy sis demonstrates that large-scale implementation of recourses by end users may e xacerbate social segregation. To address this problem, we propose novel algorith ms which leverage implicit and explicit conditional generative models to not onl y minimize the chance of segregation but also provide realistic recourses. Exten sive experimentation with real-world datasets demonstrates the efficacy of the p roposed approaches.

****

DDGR: Continual Learning with Deep Diffusion-based Generative Replay

Rui Gao, Weiwei Liu

Popular deep-learning models in the field of image classification suffer from ca tastrophic forgetting—models will forget previously acquired skills when learnin g new ones. Generative replay (GR), which typically consists of a generator and a classifier, is an efficient way to mitigate catastrophic forgetting. However, conventional GR methods only focus on a single instruction relationship (generat

or-to-classifier), where the generator synthesizes samples for previous tasks to instruct the training of the classifier, while ignoring the ways in which the classifier can benefit the generator. In addition, most generative replay methods typically reuse the generated samples to update the generator, which causes the samples regenerated by the generator deviating from the distribution of previous tasks. To overcome these two issues, we propose a novel approach, called deep diffusion-based generative replay (DDGR), which adopts a diffusion model as the generator and calculates an instruction-operator through the classifier to instruct the generation of samples. Extensive experiments in class incremental (CI) and class incremental with repetition (CIR) settings demonstrate the advantages of DDGR. Our code is available at https://github.com/xiaocangshengGR/DDGR.

****

PAL: Program-aided Language Models

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, Graham Neubig

Large language models (LLMs) have demonstrated an impressive ability to perform arithmetic and symbolic reasoning tasks, when provided with a few examples at test time ("few-shot prompting"). Much of this success can be attributed to prompting methods such as "chain-of-thought", which employ LLMs for both understanding the problem description by decomposing it into steps, as well as solving each step of the problem. While LLMs seem to be adept at this sort of step-by-step decomposition, LLMs often make logical and arithmetic mistakes in the solution part, even when the problem is decomposed correctly. In this paper, we present Program-Aided Language models (PAL): a novel approach that uses the LLM to read natural language problems and generate programs as the intermediate reasoning steps, but offloads the solution step to a runtime such as a Python interpreter. With PAL, decomposing the natural language problem into runnable steps remains the only learning task for the LLM, while solving is delegated to the interpreter. We demonstrate this synergy between a neural LLM and a symbolic interpreter across 13 mathematical, symbolic, and algorithmic reasoning tasks from BIG-Bench Hard and others. In all these natural language reasoning tasks, generating code using an LLM and reasoning using a Python interpreter leads to more accurate results than much larger models. For example, PAL using Codex achieves state-of-the-art few-shot accuracy on GSM8K, surpassing PaLM which uses chain-of-thought by absolute 15% top-1.

****

Out-of-Domain Robustness via Targeted Augmentations

Irena Gao, Shiori Sagawa, Pang Wei Koh, Tatsunori Hashimoto, Percy Liang

Models trained on one set of domains often suffer performance drops on unseen domains, e.g., when wildlife monitoring models are deployed in new camera locations. In this work, we study principles for designing data augmentations for out-of-domain (OOD) generalization. In particular, we focus on real-world scenarios in which some domain-dependent features are robust, i.e., some features that vary across domains are predictive OOD. For example, in the wildlife monitoring application above, image backgrounds vary across camera locations but indicate habitat type, which helps predict the species of photographed animals. Motivated by theoretical analysis on a linear setting, we propose targeted augmentations, which selectively randomize spurious domain-dependent features while preserving robust ones. We prove that targeted augmentations improve OOD performance, allowing models to generalize better with fewer domains. In contrast, existing approaches such as generic augmentations, which fail to randomize domain-dependent features, and domain-invariant augmentations, which randomize all domain-dependent features, both perform poorly OOD. In experiments on three real-world datasets, we show that targeted augmentations set new states-of-the-art for OOD performance by 3.2-15.2%.

****

Scaling Laws for Reward Model Overoptimization

Leo Gao, John Schulman, Jacob Hilton

In reinforcement learning from human feedback, it is common to optimize against a reward model trained to predict human preferences. Because the reward model is

an imperfect proxy, optimizing its value too much can hinder ground truth performance, in accordance with Goodhart’s law. This effect has been frequently observed, but not carefully measured due to the expense of collecting human preference data. In this work, we use a synthetic setup in which a fixed “gold-standard” reward model plays the role of humans, providing labels used to train a proxy reward model. We study how the gold reward model score changes as we optimize against the proxy reward model using either reinforcement learning or best-of-n sampling. We find that this relationship follows a different functional form depending on the method of optimization, and that in both cases its coefficients scale smoothly with the number of reward model parameters. We also study the effect on this relationship of the size of the reward model dataset, the number of reward model and policy parameters, and the coefficient of the KL penalty added to the reward in the reinforcement learning setup. We explore the implications of these empirical results for theoretical considerations in AI alignment.

****

The Unreasonable Effectiveness of Few-shot Learning for Machine Translation

Xavier Garcia, Yamini Bansal, Colin Cherry, George Foster, Maxim Krikun, Melvin Johnson, Orhan Firat

We demonstrate the potential of few-shot translation systems, trained with unpaired language data, for both high and low-resource language pairs. We show that with only 5 examples of high-quality translation data shown at inference, a transformer decoder-only model trained solely with self-supervised learning, is able to match specialized supervised state-of-the-art models as well as more general commercial translation systems. In particular, we outperform the best performing system on the WMT’21 English-Chinese news translation task by only using five examples of English-Chinese parallel data at inference. Furthermore, the resulting models are two orders of magnitude smaller than state-of-the-art language models. We then analyze the factors which impact the performance of few-shot translation systems, and highlight that the quality of the few-shot demonstrations heavily determines the quality of the translations generated by our models. Finally, we show that the few-shot paradigm also provides a way to control certain attributes of the translation — we show that we are able to control for regional varieties and formality using only a five examples at inference, paving the way towards controllable machine translation systems.

****

RLSbench: Domain Adaptation Under Relaxed Label Shift

Saurabh Garg, Nick Erickson, James Sharpnack, Alex Smola, Sivaraman Balakrishnan, Zachary Chase Lipton

Despite the emergence of principled methods for domain adaptation under label shift, their sensitivity to shifts in class conditional distributions is precariously under explored. Meanwhile, popular deep domain adaptation heuristics tend to falter when faced with label proportions shifts. While several papers modify these heuristics in attempts to handle label proportions shifts, inconsistencies in evaluation standards, datasets, and baselines make it difficult to gauge the current best practices. In this paper, we introduce RLSbench, a large-scale benchmark for relaxed label shift, consisting of $>$500 distribution shift pairs spanning vision, tabular, and language modalities, with varying label proportions. Unlike existing benchmarks, which primarily focus on shifts in class-conditional $p(x|y)$, our benchmark also focuses on label marginal shifts. First, we assess 13 popular domain adaptation methods, demonstrating more widespread failures under label proportion shifts than were previously known. Next, we develop an effective two-step meta-algorithm that is compatible with most domain adaptation heuristics: (i) pseudo-balance the data at each epoch; and (ii) adjust the final classifier with target label distribution estimate. The meta-algorithm improves existing domain adaptation heuristics under large label proportion shifts, often by 2–10% accuracy points, while conferring minimal effect ($<$0.5%) when label proportions do not shift. We hope that these findings and the availability of RLSbench will encourage researchers to rigorously evaluate proposed methods in relaxed label shift settings. Code is publicly available at https://github.com/acmi-lab/RLSbench.

****
RankMe: Assessing the Downstream Performance of Pretrained Self-Supervised Representations by Their Rank
Quentin Garrido, Randall Balestriero, Laurent Najman, Yann Lecun
Joint-Embedding Self Supervised Learning (JE-SSL) has seen a rapid development, with the emergence of many method variations but only few principled guidelines that would help practitioners to successfully deploy them. The main reason for that pitfall comes from JE-SSL's core principle of not employing any input reconstruction therefore lacking visual cues of unsuccessful training. Adding non informative loss values to that, it becomes difficult to deploy SSL on a new dataset for which no labels can help to judge the quality of the learned representation. In this study, we develop a simple unsupervised criterion that is indicative of the quality of the learned JE-SSL representations: their effective rank. Albeit simple and computationally friendly, this method —coined RankMe— allows one to assess the performance of JE-SSL representations, even on different downstream datasets, without requiring any labels. A further benefit of RankMe is that it does not have any training or hyper-parameters to tune. Through thorough empirical experiments involving hundreds of training episodes, we demonstrate how RankMe can be used for hyperparameter selection with nearly no reduction in final performance compared to the current selection method that involve a dataset's labels. We hope that RankMe will facilitate the deployment of JE-SSL towards domains that do not have the opportunity to rely on labels for representations' quality assessment.
****
Self-supervised learning of Split Invariant Equivariant representations
Quentin Garrido, Laurent Najman, Yann Lecun
Recent progress has been made towards learning invariant or equivariant representations with self-supervised learning. While invariant methods are evaluated on large scale datasets, equivariant ones are evaluated in smaller, more controlled, settings. We aim at bridging the gap between the two in order to learn more diverse representations that are suitable for a wide range of tasks. We start by introducing a dataset called 3DIEBench, consisting of renderings from 3D models over 55 classes and more than 2.5 million images where we have full control on the transformations applied to the objects. We further introduce a predictor architecture based on hypernetworks to learn equivariant representations with no possible collapse to invariance. We introduce SIE (Split Invariant-Equivariant) which combines the hypernetwork-based predictor with representations split in two parts, one invariant, the other equivariant, to learn richer representations. We demonstrate significant performance gains over existing methods on equivariance related tasks from both a qualitative and quantitative point of view. We further analyze our introduced predictor and show how it steers the learned latent space. We hope that both our introduced dataset and approach will enable learning richer representations without supervision in more complex scenarios. Code and data are available at https://github.com/garridoq/SIE.
****
Federated Heavy Hitter Recovery under Linear Sketching
Adria Gascon, Peter Kairouz, Ziteng Sun, Ananda Theertha Suresh
Motivated by real-life deployments of multi-round federated analytics with secure aggregation, we investigate the fundamental communication-accuracy tradeoffs of the heavy hitter discovery and approximate (open-domain) histogram problems under a linear sketching constraint. We propose efficient algorithms based on local subsampling and invertible bloom look-up tables (IBLTs). We also show that our algorithms are information-theoretically optimal for a broad class of interactive schemes. The results show that the linear sketching constraint does increase the communication cost for both tasks by introducing an extra linear dependence on the number of users in a round. Moreover, our results also establish a separation between the communication cost for heavy hitter discovery and approximate histogram in the multi-round setting. The dependence on the number of rounds R is at most logarithmic for heavy hitter discovery whereas that of approximate histogram is $\Theta(\sqrt{R})$. We also empirically demonstrate our findings.

****
On the Global Convergence of Fitted Q-Iteration with Two-layer Neural Network Parametrization
Mudit Gaur, Vaneet Aggarwal, Mridul Agarwal
Deep Q-learning based algorithms have been applied successfully in many decision making problems, while their theoretical foundations are not as well understood. In this paper, we study a Fitted Q-Iteration with two-layer ReLU neural network parameterization, and find the sample complexity guarantees for the algorithm. Our approach estimates the Q-function in each iteration using a convex optimization problem. We show that this approach achieves a sample complexity of $\tilde{\mathcal{O}}(1/\epsilon^{2})$, which is order-optimal. This result holds for a countable state-spaces and does not require any assumptions such as a linear or low rank structure on the MDP.
****
A Reinforcement Learning Framework for Dynamic Mediation Analysis
Lin Ge, Jitao Wang, Chengchun Shi, Zhenke Wu, Rui Song
Mediation analysis learns the causal effect transmitted via mediator variables between treatments and outcomes, and receives increasing attention in various scientific domains to elucidate causal relations. Most existing works focus on point-exposure studies where each subject only receives one treatment at a single time point. However, there are a number of applications (e.g., mobile health) where the treatments are sequentially assigned over time and the dynamic mediation effects are of primary interest. Proposing a reinforcement learning (RL) framework, we are the first to evaluate dynamic mediation effects in settings with infinite horizons. We decompose the average treatment effect into an immediate direct effect, an immediate mediation effect, a delayed direct effect, and a delayed mediation effect. Upon the identification of each effect component, we further develop robust and semi-parametrically efficient estimators under the RL framework to infer these causal effects. The superior performance of the proposed method is demonstrated through extensive numerical studies, theoretical results, and an analysis of a mobile health dataset. A Python implementation of the proposed procedure is available at https://github.com/linlinlin97/MediationRL.
****
Compositional Score Modeling for Simulation-Based Inference
Tomas Geffner, George Papamakarios, Andriy Mnih
Neural Posterior Estimation methods for simulation-based inference can be ill-suited for dealing with posterior distributions obtained by conditioning on multiple observations, as they tend to require a large number of simulator calls to learn accurate approximations. In contrast, Neural Likelihood Estimation methods can handle multiple observations at inference time after learning from individual observations, but they rely on standard inference methods, such as MCMC or variational inference, which come with certain performance drawbacks. We introduce a new method based on conditional score modeling that enjoys the benefits of both approaches. We model the scores of the (diffused) posterior distributions induced by individual observations, and introduce a way of combining the learned scores to approximately sample from the target posterior distribution. Our approach is sample-efficient, can naturally aggregate multiple observations at inference time, and avoids the drawbacks of standard inference methods.
****
Cramming: Training a Language Model on a single GPU in one day.
Jonas Geiping, Tom Goldstein
Recent trends in language modeling have focused on increasing performance through scaling, and have resulted in an environment where training language models is out of reach for most researchers and practitioners. While most in the community are asking how to push the limits of extreme computation, we ask the opposite question: How far can we get with a single GPU in just one day? We investigate the downstream performance achievable with a transformer-based language model trained completely from scratch with masked language modeling for a single day on a single consumer GPU. Aside from re-analyzing nearly all components of the pretraining pipeline for this scenario and providing a modified pipeline with perform

ance close to BERT, we investigate why scaling down is hard, and which modificat ions actually improve performance in this scenario. We provide evidence that eve n in this constrained setting, performance closely follows scaling laws observed in large-compute settings. Through the lens of scaling laws, we categorize a ra nge of recent improvements to training and architecture and discuss their merit and practical applicability (or lack thereof) for the limited compute setting. W e provide code to reproduce all experiments at github.com/JonasGeiping/cramming .

****

Transformers Meet Directed Graphs

Simon Geisler, Yujia Li, Daniel J Mankowitz, Ali Taylan Cemgil, Stephan Günneman n, Cosmin Paduraru

Transformers were originally proposed as a sequence-to-sequence model for text b ut have become vital for a wide range of modalities, including images, audio, vi deo, and undirected graphs. However, transformers for directed graphs are a surp risingly underexplored topic, despite their applicability to ubiquitous domains, including source code and logic circuits. In this work, we propose two directio n- and structure-aware positional encodings for directed graphs: (1) the eigenve ctors of the Magnetic Laplacian — a direction-aware generalization of the combin atorial Laplacian; (2) directional random walk encodings. Empirically, we show t hat the extra directionality information is useful in various downstream tasks, including correctness testing of sorting networks and source code understanding. Together with a data-flow-centric graph construction, our model outperforms the prior state of the art on the Open Graph Benchmark Code2 relatively by 14.7%.

****

Memory-Based Meta-Learning on Non-Stationary Distributions

Tim Genewein, Gregoire Deletang, Anian Ruoss, Li Kevin Wenliang, Elliot Catt, Vi ncent Dutordoir, Jordi Grau-Moya, Laurent Orseau, Marcus Hutter, Joel Veness

Memory-based meta-learning is a technique for approximating Bayes-optimal predic tors. Under fairly general conditions, minimizing sequential prediction error, m easured by the log loss, leads to implicit meta-learning. The goal of this work is to investigate how far this interpretation can be realized by current sequenc e prediction models and training regimes. The focus is on piecewise stationary s ources with unobserved switching-points, which arguably capture an important cha racteristic of natural language and action-observation sequences in partially ob servable environments. We show that various types of memory-based neural models, including Transformers, LSTMs, and RNNs can learn to accurately approximate kno wn Bayes-optimal algorithms and behave as if performing Bayesian inference over the latent switching-points and the latent parameters governing the data distrib ution within each segment.

****

Towards Reliable Neural Specifications

Chuqin Geng, Nham Le, Xiaojie Xu, Zhaoyue Wang, Arie Gurfinkel, Xujie Si

Having reliable specifications is an unavoidable challenge in achieving verifiab le correctness, robustness, and interpretability of AI systems. Existing specifi cations for neural networks are in the paradigm of data as specification. That i s, the local neighborhood centering around a reference input is considered to be correct (or robust). While existing specifications contribute to verifying adve rsarial robustness, a significant problem in many research domains, our empirica l study shows that those verified regions are somewhat tight, and thus fail to a llow verification of test set inputs, making them impractical for some real-worl d applications. To this end, we propose a new family of specifications called ne ural representation as specification. This form of specifications uses the intri nsic information of neural networks, specifically neural activation patterns (NA Ps), rather than input data to specify the correctness and/or robustness of neur al network predictions. We present a simple statistical approach to mining neura l activation patterns. To show the effectiveness of discovered NAPs, we formally verify several important properties, such as various types of misclassification s will never happen for a given NAP, and there is no ambiguity between different NAPs. We show that by using NAP, we can verify a significant region of the inpu

t space, while still recalling 84% of the data on MNIST. Moreover, we can push t he verifiable bound to 10 times larger on the CIFAR10 benchmark. Thus, we argue that NAPs can potentially be used as a more reliable and extensible specificatio n for neural network verification.
****

Oracles & Followers: Stackelberg Equilibria in Deep Multi-Agent Reinforcement Le arning
Matthias Gerstgrasser, David C. Parkes
Stackelberg equilibria arise naturally in a range of popular learning problems, such as in security games or indirect mechanism design, and have received increa sing attention in the reinforcement learning literature. We present a general fr amework for implementing Stackelberg equilibria search as a multi-agent RL probl em, allowing a wide range of algorithmic design choices. We discuss how previous approaches can be seen as specific instantiations of this framework. As a key i nsight, we note that the design space allows for approaches not previously seen in the literature, for instance by leveraging multitask and meta-RL techniques f or follower convergence. We propose one such approach using contextual policies, and evaluate it experimentally on both standard and novel benchmark domains, sh owing greatly improved sample efficiency compared to previous approaches. Finall y, we explore the effect of adopting algorithm designs outside the borders of ou r framework.
****

Approximately Optimal Core Shapes for Tensor Decompositions
Mehrdad Ghadiri, Matthew Fahrbach, Gang Fu, Vahab Mirrokni
This work studies the combinatorial optimization problem of finding an optimal c ore tensor shape, also called multilinear rank, for a size-constrained Tucker de composition. We give an algorithm with provable approximation guarantees for its reconstruction error via connections to higher-order singular values. Specifica lly, we introduce a novel Tucker packing problem, which we prove is NP-hard, and give a polynomial-time approximation scheme based on a reduction to the 2-dimen sional knapsack problem with a matroid constraint. We also generalize our techni ques to tree tensor network decompositions. We implement our algorithm using an integer programming solver, and show that its solution quality is competitive wi th (and sometimes better than) the greedy algorithm that uses the true Tucker de composition loss at each step, while also running up to 1000x faster.
****

GAT: Guided Adversarial Training with Pareto-optimal Auxiliary Tasks
Salah Ghamizi, Jingfeng Zhang, Maxime Cordy, Mike Papadakis, Masashi Sugiyama, Y ves Le Traon
While leveraging additional training data is well established to improve adversa rial robustness, it incurs the unavoidable cost of data collection and the heavy computation to train models. To mitigate the costs, we propose *Guided Adversar ial Training * (GAT), a novel adversarial training technique that exploits auxil iary tasks under a limited set of training data. Our approach extends single-tas k models into multi-task models during the min-max optimization of adversarial t raining, and drives the loss optimization with a regularization of the gradient curvature across multiple tasks. GAT leverages two types of auxiliary tasks: sel f-supervised tasks, where the labels are generated automatically, and domain-kno wledge tasks, where human experts provide additional labels. Experimentally, und er limited data, GAT increases the robust accuracy on CIFAR-10 up to four times (from 11% to 42% robust accuracy) and the robust AUC of CheXpert medical imaging dataset from 50% to 83%. On the full CIFAR-10 dataset, GAT outperforms eight st ate-of-the-art adversarial training strategies. Our large study across five data sets and six tasks demonstrates that task augmentation is an efficient alternati ve to data augmentation, and can be key to achieving both clean and robust perfo rmances.
****

On User-Level Private Convex Optimization
Badih Ghazi, Pritish Kamath, Ravi Kumar, Pasin Manurangsi, Raghu Meka, Chiyuan Z hang

We introduce a new mechanism for stochastic convex optimization (SCO) with user-level differential privacy guarantees. The convergence rates of this mechanism are similar to those in the prior work of Levy et al. 2021 and Narayanan et al. 2022, but with two important improvements. Our mechanism does not require any smoothness assumptions on the loss. Furthermore, our bounds are also the first where the minimum number of users needed for user-level privacy has no dependence on the dimension and only a logarithmic dependence on the desired excess error. The main idea underlying the new mechanism is to show that the optimizers of strongly convex losses have low local deletion sensitivity, along with a new output perturbation method for functions with low local deletion sensitivity, which could be of independent interest.
****

Contextual Reliability: When Different Features Matter in Different Contexts
Gaurav Rohit Ghosal, Amrith Setlur, Daniel S. Brown, Anca Dragan, Aditi Raghunathan

Deep neural networks often fail catastrophically by relying on spurious correlations. Most prior work assumes a clear dichotomy into spurious and reliable features; however, this is often unrealistic. For example, most of the time we do not want an autonomous car to simply copy the speed of surrounding cars—we don't want our car to run a red light if a neighboring car does so. However, we cannot simply enforce invariance to next-lane speed, since it could provide valuable information about an unobservable pedestrian at a crosswalk. Thus, universally ignoring features that are sometimes (but not always) reliable can lead to non-robust performance. We formalize a new setting called contextual reliability which accounts for the fact that the "right" features to use may vary depending on the context. We propose and analyze a two-stage framework called Explicit Non-spurious feature Prediction (ENP) which first identifies the relevant features to use for a given context, then trains a model to rely exclusively on these features. Our work theoretically and empirically demonstrates the advantages of ENP over existing methods and provides new benchmarks for contextual reliability.
****

Reinforcement Learning from Passive Data via Latent Intentions
Dibya Ghosh, Chethan Anand Bhateja, Sergey Levine

Passive observational data, such as human videos, is abundant and rich in information, yet remains largely untapped by current RL methods. Perhaps surprisingly, we show that passive data, despite not having reward or action labels, can still be used to learn features that accelerate downstream RL. Our approach learns from passive data by modeling intentions: measuring how the likelihood of future outcomes change when the agent acts to achieve a particular task. We propose a temporal difference learning objective to learn about intentions, resulting in an algorithm similar to conventional RL, but which learns entirely from passive data. When optimizing this objective, our agent simultaneously learns representations of states, of policies, and of possible outcomes in an environment, all from raw observational data. Both theoretically and empirically, this scheme learns features amenable for value prediction for downstream tasks, and our experiments demonstrate the ability to learn from many forms of passive data, including cross-embodiment video data and YouTube videos.
****

Harmonic Neural Networks
Atiyo Ghosh, Antonio Andrea Gentile, Mario Dagrada, Chul Lee, Seong-Hyok Sean Kim, Hyukgeun Cha, Yunjun Choi, Dongho Kim, Jeong-Il Kye, Vincent Emanuel Elfving

Harmonic functions are abundant in nature, appearing in limiting cases of Maxwell's, Navier-Stokes equations, the heat and the wave equation. Consequently, there are many applications of harmonic functions from industrial process optimisation to robotic path planning and the calculation of first exit times of random walks. Despite their ubiquity and relevance, there have been few attempts to incorporate inductive biases towards harmonic functions in machine learning contexts. In this work, we demonstrate effective means of representing harmonic functions in neural networks and extend such results also to quantum neural networks to demonstrate the generality of our approach. We benchmark our approaches against (

quantum) physics-informed neural networks, where we show favourable performance.
****

Dividing and Conquering a BlackBox to a Mixture of Interpretable Models: Route, Interpret, Repeat
Shantanu Ghosh, Ke Yu, Forough Arabshahi, Kayhan Batmanghelich
ML model design either starts with an interpretable model or a Blackbox and explains it post hoc. Blackbox models are flexible but difficult to explain, while interpretable models are inherently explainable. Yet, interpretable models require extensive ML knowledge and tend to be less flexible, potentially underperforming than their Blackbox equivalents. This paper aims to blur the distinction between a post hoc explanation of a Blackbox and constructing interpretable models. Beginning with a Blackbox, we iteratively carve out a mixture of interpretable models and a residual network. The interpretable models identify a subset of samples and explain them using First Order Logic (FOL), providing basic reasoning on concepts from the Blackbox. We route the remaining samples through a flexible residual. We repeat the method on the residual network until all the interpretable models explain the desired proportion of data. Our extensive experiments show that our route, interpret, and repeat approach (1) identifies a richer diverse set of instance-specific concepts with high concept completeness via interpretable models by specializing in various subsets of data without compromising in performance, (2) identifies the relatively “harder” samples to explain via residuals, (3) outperforms the interpretable by-design models by significant margins during test-time interventions, (4) can be used to fix the shortcut learned by the original Blackbox.
****

Looped Transformers as Programmable Computers
Angeliki Giannou, Shashank Rajput, Jy-Yong Sohn, Kangwook Lee, Jason D. Lee, Dimitris Papailiopoulos
We present a framework for using transformer networks as universal computers by programming them with specific weights and placing them in a loop. Our input sequence acts as a punchcard, consisting of instructions and memory for data read/writes. We demonstrate that a constant number of encoder layers can emulate basic computing blocks, including lexicographic operations, non-linear functions, function calls, program counters, and conditional branches. Using this framework, we emulate a computer using a simple instruction-set architecture, which allows us to map iterative algorithms to programs that can be executed by a constant depth looped transformer network. We show how a single frozen transformer, instructed by its input, can emulate a basic calculator, a basic linear algebra library, and even a full backpropagation, in-context learning algorithm. Our findings reveal the potential of transformer networks as programmable compute units and offer insight into the mechanics of attention.
****

Generalized Disparate Impact for Configurable Fairness Solutions in ML
Luca Giuliani, Eleonora Misino, Michele Lombardi
We make two contributions in the field of AI fairness over continuous protected attributes. First, we show that the Hirschfeld-Gebelein-Renyi (HGR) indicator (the only one currently available for such a case) is valuable but subject to a few crucial limitations regarding semantics, interpretability, and robustness. Second, we introduce a family of indicators that are: 1) complementary to HGR in terms of semantics; 2) fully interpretable and transparent; 3) robust over finite samples; 4) configurable to suit specific applications. Our approach also allows us to define fine-grained constraints to permit certain types of dependence and forbid others selectively. By expanding the available options for continuous protected attributes, our approach represents a significant contribution to the area of fair artificial intelligence.
****

Multicalibration as Boosting for Regression
Ira Globus-Harris, Declan Harrison, Michael Kearns, Aaron Roth, Jessica Sorrell
We study the connection between multicalibration and boosting for squared error regression. First we prove a useful characterization of multicalibration in term

s of a “swap regret” like condition on squared error. Using this characterizatio n, we give an exceedingly simple algorithm that can be analyzed both as a boosti ng algorithm for regression and as a multicalibration algorithm for a class $\ma thcal{H}$ that makes use only of a standard squared error regression oracle for $\mathcal{H}$. We give a weak learning assumption on $\mathcal{H}$ that ensures convergence to Bayes optimality without the need to make any realizability assum ptions — giving us an agnostic boosting algorithm for regression. We then show t hat our weak learning assumption on $\mathcal{H}$ is both necessary and sufficie nt for multicalibration with respect to $\mathcal{H}$ to imply Bayes optimality, answering an open question. We also show that if $\mathcal{H}$ satisfies our we ak learning condition relative to another class $\mathcal{C}$ then multicalibrat ion with respect to $\mathcal{H}$ implies multicalibration with respect to $\mat hcal{C}$. Finally we investigate the empirical performance of our algorithm expe rimentally.

****

Adversarial robustness of amortized Bayesian inference

Manuel Gloeckler, Michael Deistler, Jakob H. Macke

Bayesian inference usually requires running potentially costly inference procedu res separately for every new observation. In contrast, the idea of amortized Bay esian inference is to initially invest computational cost in training an inferen ce network on simulated data, which can subsequently be used to rapidly perform inference (i.e., to return estimates of posterior distributions) for new observa tions. This approach has been applied to many real-world models in the sciences and engineering, but it is unclear how robust the approach is to adversarial per turbations in the observed data. Here, we study the adversarial robustness of am ortized Bayesian inference, focusing on simulation-based estimation of multi-dim ensional posterior distributions. We show that almost unrecognizable, targeted p erturbations of the observations can lead to drastic changes in the predicted po sterior and highly unrealistic posterior predictive samples, across several benc hmark tasks and a real-world example from neuroscience. We propose a computation ally efficient regularization scheme based on penalizing the Fisher information of the conditional density estimator, and show how it improves the adversarial r obustness of amortized Bayesian inference.

****

Efficient RL via Disentangled Environment and Agent Representations

Kevin Gmelin, Shikhar Bahl, Russell Mendonca, Deepak Pathak

Agents that are aware of the separation between the environments and themselves can leverage this understanding to form effective representations of visual inpu t. We propose an approach for learning such structured representations for RL al gorithms, using visual knowledge of the agent, which is often inexpensive to obt ain, such as its shape or mask. This is incorporated into the RL objective using a simple auxiliary loss. We show that our method, SEAR (Structured Environment-Agent Representations), outperforms state-of-the-art model-free approaches over 18 different challenging visual simulation environments spanning 5 different rob ots.

****

Aligning Language Models with Preferences through f-divergence Minimization

Dongyoung Go, Tomasz Korbak, Germàn Kruszewski, Jos Rozen, Nahyeon Ryu, Marc Dym etman

Aligning language models with preferences can be posed as approximating a target distribution representing some desired behavior. Existing approaches differ bot h in the functional form of the target distribution and the algorithm used to ap proximate it. For instance, Reinforcement Learning from Human Feedback (RLHF) co rresponds to minimizing a reverse KL from an implicit target distribution arisin g from a KL penalty in the objective. On the other hand, Generative Distribution al Control (GDC) has an explicit target distribution and minimizes a forward KL from it using the Distributional Policy Gradient (DPG) algorithm. In this paper, we propose a new approach, f-DPG, which allows the use of any f-divergence to approximate any target distribution that can be evaluated. f-DPG unifies bo th frameworks (RLHF, GDC) and the approximation methods (DPG, RL with KL penalti

es). We show the practical benefits of various choices of divergence objectives and demonstrate that there is no universally optimal objective but that different divergences present different alignment and diversity trade-offs. We show that Jensen-Shannon divergence strikes a good balance between these objectives, and frequently outperforms forward KL divergence by a wide margin, leading to significant improvements over prior work. These distinguishing characteristics between divergences persist as the model size increases, highlighting the importance of selecting appropriate divergence objectives.

****

Robust Consensus in Ranking Data Analysis: Definitions, Properties and Computational Issues

Morgane Goibert, Clément Calauzènes, Ekhine Irurozki, Stephan Clémençon

As the issue of robustness in AI systems becomes vital, statistical learning techniques that are reliable even in presence of partly contaminated data have to be developed. Preference data, in the form of (complete) rankings in the simplest situations, are no exception and the demand for appropriate concepts and tools is all the more pressing given that technologies fed by or producing this type of data ($\textit{e.g.}$ search engines, recommending systems) are now massively deployed. However, the lack of vector space structure for the set of rankings ($\textit{i.e.}$ the symmetric group $\mathfrak{S}_n$) and the complex nature of statistics considered in ranking data analysis make the formulation of robustness objectives in this domain challenging. In this paper, we introduce notions of robustness, together with dedicated statistical methods, for $\textit{Consensus Ranking}$ the flagship problem in ranking data analysis, aiming at summarizing a probability distribution on $\mathfrak{S}_n$ by a $\textit{median}$ ranking. Precisely, we propose specific extensions of the popular concept of breakdown point, tailored to consensus ranking, and address the related computational issues. Beyond the theoretical contributions, the relevance of the approach proposed is supported by an experimental study.

****

Learning Distributions over Quantum Measurement Outcomes

Weiyuan Gong, Scott Aaronson

Shadow tomography for quantum states provides a sample efficient approach for predicting the measurement outcomes of quantum systems. However, these shadow tomography procedures yield poor bounds if there are more than two outcomes per measurement. In this paper, we consider a general problem of learning properties from quantum states: given an unknown d-dimensional quantum state ρ and M unknown quantum measurements $\mathcal{M}_1,...,\mathcal{M}_M$ with $K\geq 2$ outcomes, estimating the probability distribution for applying $\mathcal{M}_i$ on ρ to within total variation distance ϵ. Compared to the special case when $K=2$, we have to learn unknown distributions instead of values. Here, we propose an online shadow tomography procedure that solves this problem with high success probability requiring $\tilde{O}(K\log^2M\log d/\epsilon^4)$ copies of ρ. We further prove an information-theoretic lower bound showing that at least $\Omega(\min\{d^2,K+\log M\}/\epsilon^2)$ copies of ρ are required to solve this problem with high success probability. Our shadow tomography procedure requires sample complexity with only logarithmic dependence on M and d and is sample-optimal concerning the dependence on K.

****

Convergence of Proximal Point and Extragradient-Based Methods Beyond Monotonicity: the Case of Negative Comonotonicity

Eduard Gorbunov, Adrien Taylor, Samuel Horváth, Gauthier Gidel

Algorithms for min-max optimization and variational inequalities are often studied under monotonicity assumptions. Motivated by non-monotone machine learning applications, we follow the line of works (Diakonikolas et al., 2021; Lee & Kim, 2021; Pethick et al., 2022; Bohm,2022) aiming at going beyond monotonicity by considering the weaker negative comonotonicity assumption. In this work, we provide tight complexity analyses for the Proximal Point (PP), Extragradient (EG), and Optimistic Gradient (OG) methods in this setup, closing several questions on their working guarantees beyond monotonicity. In particular, we derive the first no

n-asymptotic convergence rates for PP under negative comonotonicity and star-neg
ative comonotonicity and show their tightness via constructing worst-case exampl
es; we also relax the assumptions for the last-iterate convergence guarantees fo
r EG and OG and prove the tightness of the existing best-iterate guarantees for
EG and OG via constructing counter-examples.
****

Adaptive Annealed Importance Sampling with Constant Rate Progress
Shirin Goshtasbpour, Victor Cohen, Fernando Perez-Cruz
Annealed Importance Sampling (AIS) synthesizes weighted samples from an intracta
ble distribution given its unnormalized density function. This algorithm relies
on a sequence of interpolating distributions bridging the target to an initial t
ractable distribution such as the well-known geometric mean path of unnormalized
 distributions which is assumed to be suboptimal in general. In this paper, we p
rove that the geometric annealing corresponds to the distribution path that mini
mizes the KL divergence between the current particle distribution and the desire
d target when the feasible change in the particle distribution is constrained. F
ollowing this observation, we derive the constant rate discretization schedule f
or this annealing sequence, which adjusts the schedule to the difficulty of movi
ng samples between the initial and the target distributions. We further extend o
ur results to f-divergences and present the respective dynamics of annealing s
equences based on which we propose the Constant Rate AIS (CR-AIS) algorithm and
its efficient implementation for α-divergences. We empirically show that
CR-AIS performs well on multiple benchmark distributions while avoiding the comp
utationally expensive tuning loop in existing Adaptive AIS.
****

Formalizing Preferences Over Runtime Distributions
Devon R. Graham, Kevin Leyton-Brown, Tim Roughgarden
When trying to solve a computational problem, we are often faced with a choice b
etween algorithms that are guaranteed to return the right answer but differ in t
heir runtime distributions (e.g., SAT solvers, sorting algorithms). This paper a
ims to lay theoretical foundations for such choices by formalizing preferences o
ver runtime distributions. It might seem that we should simply prefer the algori
thm that minimizes expected runtime. However, such preferences would be driven b
y exactly how slow our algorithm is on bad inputs, whereas in practice we are ty
pically willing to cut off occasional, sufficiently long runs before they finish
. We propose a principled alternative, taking a utility-theoretic approach to ch
aracterize the scoring functions that describe preferences over algorithms. Thes
e functions depend on the way our value for solving our problem decreases with t
ime and on the distribution from which captimes are drawn. We describe examples
of realistic utility functions and show how to leverage a maximum-entropy approa
ch for modeling underspecified captime distributions. Finally, we show how to ef
ficiently estimate an algorithm’s expected utility from runtime samples.
****

Topological Point Cloud Clustering
Vincent Peter Grande, Michael T Schaub
We present Topological Point Cloud Clustering (TPCC), a new method to cluster po
ints in an arbitrary point cloud based on their contribution to global topologic
al features. TPCC synthesizes desirable features from spectral clustering and to
pological data analysis and is based on considering the spectral properties of a
 simplicial complex associated to the considered point cloud. As it is based on
considering sparse eigenvector computations, TPCC is similarly easy to interpret
 and implement as spectral clustering. However, by focusing not just on a single
 matrix associated to a graph created from the point cloud data, but on a whole
set of Hodge-Laplacians associated to an appropriately constructed simplicial co
mplex, we can leverage a far richer set of topological features to characterize
the data points within the point cloud and benefit from the relative robustness
of topological techniques against noise. We test the performance of TPCC on both
 synthetic and real-world data and compare it with classical spectral clustering
.
****

On Sampling with Approximate Transport Maps
Louis Grenioux, Alain Oliviero Durmus, Eric Moulines, Marylou Gabrié
Transport maps can ease the sampling of distributions with non-trivial geometrie
s by transforming them into distributions that are easier to handle. The potenti
al of this approach has risen with the development of Normalizing Flows (NF) whi
ch are maps parameterized with deep neural networks trained to push a reference
distribution towards a target. NF-enhanced samplers recently proposed blend (Mar
kov chain) Monte Carlo methods with either (i) proposal draws from the flow or (
ii) a flow-based reparametrization. In both cases, the quality of the learned tr
ansport conditions performance. The present work clarifies for the first time th
e relative strengths and weaknesses of these two approaches. Our study concludes
 that multimodal targets can be reliably handled with flow-based proposals up to
 moderately high dimensions. In contrast, methods relying on reparametrization s
truggle with multimodality but are more robust otherwise in high-dimensional set
tings and under poor training. To further illustrate the influence of target-pro
posal adequacy, we also derive a new quantitative bound for the mixing time of t
he Independent Metropolis-Hastings sampler.
****

Hidden Symmetries of ReLU Networks
Elisenda Grigsby, Kathryn Lindsey, David Rolnick
The parameter space for any fixed architecture of feedforward ReLU neural networ
ks serves as a proxy during training for the associated class of functions - but
 how faithful is this representation? It is known that many different parameter
settings θ can determine the same function f. Moreover, the degree of t
his redundancy is inhomogeneous: for some networks, the only symmetries are perm
utation of neurons in a layer and positive scaling of parameters at a neuron, wh
ile other networks admit additional hidden symmetries. In this work, we prove th
at, for any network architecture where no layer is narrower than the input, ther
e exist parameter settings with no hidden symmetries. We also describe a number
of mechanisms through which hidden symmetries can arise, and empirically approxi
mate the functional dimension of different network architectures at initializati
on. These experiments indicate that the probability that a network has no hidden
 symmetries decreases towards 0 as depth increases, while increasing towards 1 a
s width and input dimension increase.
****

EF21-P and Friends: Improved Theoretical Communication Complexity for Distribute
d Optimization with Bidirectional Compression
Kaja Gruntkowska, Alexander Tyurin, Peter Richtárik
In this work we focus our attention on distributed optimization problems in the
context where the communication time between the server and the workers is non-n
egligible. We obtain novel methods supporting bidirectional compression (both fr
om the server to the workers and vice versa) that enjoy new state-of-the-art the
oretical communication complexity for convex and nonconvex problems. Our bounds
are the first that manage to decouple the variance/error coming from the workers
-to-server and server-to-workers compression, transforming a multiplicative depe
ndence to an additive one. Moreover, in the convex regime, we obtain the first b
ounds that match the theoretical communication complexity of gradient descent. E
ven in this convex regime, our algorithms work with biased gradient estimators,
which is non-standard and requires new proof techniques that may be of independe
nt interest. Finally, our theoretical results are corroborated through suitable
experiments.
****

NerfDiff: Single-image View Synthesis with NeRF-guided Distillation from 3D-awar
e Diffusion
Jiatao Gu, Alex Trevithick, Kai-En Lin, Joshua M. Susskind, Christian Theobalt,
Lingjie Liu, Ravi Ramamoorthi
Novel view synthesis from a single image requires inferring occluded regions of
objects and scenes whilst simultaneously maintaining semantic and physical consi
stency with the input. Existing approaches condition neural radiance fields (NeR
F) on local image features, projecting points to the input image plane, and aggr

egating 2D features to perform volume rendering. However, under severe occlusion , this projection fails to resolve uncertainty, resulting in blurry renderings t hat lack details. In this work, we propose NerfDiff, which addresses this issue by distilling the knowledge of a 3D-aware conditional diffusion model (CDM) into NeRF through synthesizing and refining a set of virtual views at test-time. We further propose a novel NeRF-guided distillation algorithm that simultaneously g enerates 3D consistent virtual views from the CDM samples, and finetunes the NeR F based on the improved virtual views. Our approach significantly outperforms ex isting NeRF-based and geometry-free approaches on challenging datasets including ShapeNet, ABO, and Clevr3D.

****

DecompDiff: Diffusion Models with Decomposed Priors for Structure-Based Drug Des ign

Jiaqi Guan, Xiangxin Zhou, Yuwei Yang, Yu Bao, Jian Peng, Jianzhu Ma, Qiang Liu, Liang Wang, Quanquan Gu

Designing 3D ligands within a target binding site is a fundamental task in drug discovery. Existing structured-based drug design methods treat all ligand atoms equally, which ignores different roles of atoms in the ligand for drug design an d can be less efficient for exploring the large drug-like molecule space. In thi s paper, inspired by the convention in pharmaceutical practice, we decompose the ligand molecule into two parts, namely arms and scaffold, and propose a new dif fusion model, DecompDiff, with decomposed priors over arms and scaffold. In orde r to facilitate the decomposed generation and improve the properties of the gene rated molecules, we incorporate both bond diffusion in the model and additional validity guidance in the sampling phase. Extensive experiments on CrossDocked202 0 show that our approach achieves state-of-the-art performance in generating hig h-affinity molecules while maintaining proper molecular properties and conformat ional stability, with up to -8.39 Avg. Vina Dock score and $24.5%$ Success Rat e. The code is provided at https://github.com/bytedance/DecompDiff

****

On Excess Mass Behavior in Gaussian Mixture Models with Orlicz-Wasserstein Dista nces

Aritra Guha, Nhat Ho, Xuanlong Nguyen

Dirichlet Process mixture models (DPMM) in combination with Gaussian kernels hav e been an important modeling tool for numerous data domains arising from biologi cal, physical, and social sciences. However, this versatility in applications do es not extend to strong theoretical guarantees for the underlying parameter esti mates, for which only a logarithmic rate is achieved. In this work, we (re)intro duce and investigate a metric, named Orlicz-Wasserstein distance, in the study o f the Bayesian contraction behavior for the parameters. We show that despite the overall slow convergence guarantees for all the parameters, posterior contracti on for parameters happens at almost polynomial rates in outlier regions of the p arameter space. Our theoretical results provide new insight in understanding the convergence behavior of parameters arising from various settings of hierarchica l Bayesian nonparametric models. In addition, we provide an algorithm to compute the metric by leveraging Sinkhorn divergences and validate our findings through a simulation study.

****

Conformalization of Sparse Generalized Linear Models

Etash Kumar Guha, Eugene Ndiaye, Xiaoming Huo

Given a sequence of observable variables $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, t he conformal prediction method estimates a confidence set for y_{n+1} given $x _{n+1}$ that is valid for any finite sample size by merely assuming that the joi nt distribution of the data is permutation invariant. Although attractive, compu ting such a set is computationally infeasible in most regression problems. Indee d, in these cases, the unknown variable y_{n+1} can take an infinite number of possible candidate values, and generating conformal sets requires retraining a predictive model for each candidate. In this paper, we focus on a sparse linear model with only a subset of variables for prediction and use numerical continuat ion techniques to approximate the solution path efficiently. The critical proper

ty we exploit is that the set of selected variables is invariant under a small p
erturbation of the input data. Therefore, it is sufficient to enumerate and refi
t the model only at the change points of the set of active features and smoothly
 interpolate the rest of the solution via a Predictor-Corrector mechanism. We sh
ow how our path-following algorithm accurately approximates conformal prediction
 sets and illustrate its performance using synthetic and real data examples.
****

Privacy-Aware Compression for Federated Learning Through Numerical Mechanism Des
ign

Chuan Guo, Kamalika Chaudhuri, Pierre Stock, Michael Rabbat

In private federated learning (FL), a server aggregates differentially private u
pdates from a large number of clients in order to train a machine learning model
. The main challenge in this setting is balancing privacy with both classificati
on accuracy of the learnt model as well as the number of bits communicated betwe
en the clients and server. Prior work has achieved a good trade-off by designing
 a privacy-aware compression mechanism, called the minimum variance unbiased (MV
U) mechanism, that numerically solves an optimization problem to determine the p
arameters of the mechanism. This paper builds upon it by introducing a new inter
polation procedure in the numerical design process that allows for a far more ef
ficient privacy analysis. The result is the new Interpolated MVU mechanism that
is more scalable, has a better privacy-utility trade-off, and provides SOTA resu
lts on communication-efficient private FL on a variety of datasets.
****

Out-of-Distribution Generalization of Federated Learning via Implicit Invariant
Relationships

Yaming Guo, Kai Guo, Xiaofeng Cao, Tieru Wu, Yi Chang

Out-of-distribution generalization is challenging for non-participating clients
of federated learning under distribution shifts. A proven strategy is to explore
 those invariant relationships between input and target variables, working equal
ly well for non-participating clients. However, learning invariant relationships
 is often in an explicit manner from data, representation, and distribution, whi
ch violates the federated principles of privacy-preserving and limited communica
tion. In this paper, we propose FedIIR, which implicitly learns invariant relati
onships from parameter for out-of-distribution generalization, adhering to the a
bove principles. Specifically, we utilize the prediction disagreement to quantif
y invariant relationships and implicitly reduce it through inter-client gradient
 alignment. Theoretically, we demonstrate the range of non-participating clients
 to which FedIIR is expected to generalize and present the convergence results f
or FedIIR in the massively distributed with limited communication. Extensive exp
eriments show that FedIIR significantly outperforms relevant baselines in terms
of out-of-distribution generalization of federated learning.
****

FeDXL: Provable Federated Learning for Deep X-Risk Optimization

Zhishuai Guo, Rong Jin, Jiebo Luo, Tianbao Yang

In this paper, we tackle a novel federated learning (FL) problem for optimizing
 a family of X-risks, to which no existing FL algorithms are applicable. In parti
cular, the objective has the form of $\mathbb{E}_{\mathbf{z}\sim \mathcal{S}_1}
f(\mathbb{E}_{\mathbf{z}'\sim\mathcal{S}_2} \ell(\mathbf{w}; \mathbf{z}, \mathbf
{z}'))$, where two sets of data $\mathcal S_1, \mathcal S_2$ are distributed ove
r multiple machines, $\ell(\cdot; \cdot,\cdot)$ is a pairwise loss that only dep
ends on the prediction outputs of the input data pairs $(\mathbf{z}, \mathbf{z}'
)$. This problem has important applications in machine learning, e.g., AUROC max
imization with a pairwise loss, and partial AUROC maximization with a compositio
nal loss. The challenges for designing an FL algorithm for X-risks lie in the no
n-decomposability of the objective over multiple machines and the interdependenc
y between different machines. To this end, we propose an active-passive decompos
ition framework that decouples the gradient’s components with two types, namely
 active parts and passive parts, where the active parts depend on local data that
 are computed with the local model and the passive parts depend on other machine
s that are communicated/computed based on historical models and samples. Under t

his framework, we design two FL algorithms (FeDXL) for handling linear and nonli
near f, respectively, based on federated averaging and merging and develop a n
ovel theoretical analysis to combat the latency of the passive parts and the int
erdependency between the local model parameters and the involved data for comput
ing local gradient estimators. We establish both iteration and communication com
plexities and show that using the historical samples and models for computing th
e passive parts do not degrade the complexities. We conduct empirical studies of
 FeDXL for deep AUROC and partial AUROC maximization, and demonstrate their perf
ormance compared with several baselines.
****
Provably Efficient Representation Learning with Tractable Planning in Low-Rank P
OMDP
Jiacheng Guo, Zihao Li, Huazheng Wang, Mengdi Wang, Zhuoran Yang, Xuezhou Zhang
In this paper, we study representation learning in partially observable Markov D
ecision Processes (POMDPs), where the agent learns a decoder function that maps
a series of high-dimensional raw observations to a compact representation and us
es it for more efficient exploration and planning. We focus our attention on the
 sub-classes of γ-observable and decodable POMDPs, for which it has been
shown that statistically tractable learning is possible, but there has not been
any computationally efficient algorithm. We first present an algorithm for decod
able PMMDPs that combines maximum likelihood estimation (MLE) and optimism in th
e face of uncertainty (OFU) to perform representation learning and achieve effic
ient sample complexity, while only calling supervised learning computational ora
cles. We then show how to adapt this algorithm to also work in the broader class
 of γ-observable POMDPs.
****
Analyzing Privacy Leakage in Machine Learning via Multiple Hypothesis Testing: A
 Lesson From Fano
Chuan Guo, Alexandre Sablayrolles, Maziar Sanjabi
Differential privacy (DP) is by far the most widely accepted framework for mitig
ating privacy risks in machine learning. However, exactly how small the privacy
parameter ϵ needs to be to protect against certain privacy risks in pra
ctice is still not well-understood. In this work, we study data reconstruction a
ttacks for discrete data and analyze it under the framework of multiple hypothes
is testing. For a learning algorithm satisfying (α, ϵ)-Renyi DP, w
e utilize different variants of the celebrated Fano’s inequality to upper bound
the attack advantage of a data reconstruction adversary. Our bound can be numeri
cally computed to relate the parameter ϵ to the desired level of privac
y protection in practice, and complements the empirical evidence for the effecti
veness of DP against data reconstruction attacks even at relatively large values
 of ϵ.
****
Linkless Link Prediction via Relational Distillation
Zhichun Guo, William Shiao, Shichang Zhang, Yozen Liu, Nitesh V Chawla, Neil Sha
h, Tong Zhao
Graph Neural Networks (GNNs) have shown exceptional performance in the task of l
ink prediction. Despite their effectiveness, the high latency brought by non-tri
vial neighborhood data dependency limits GNNs in practical deployments. Converse
ly, the known efficient MLPs are much less effective than GNNs due to the lack o
f relational knowledge. In this work, to combine the advantages of GNNs and MLPs
, we start with exploring direct knowledge distillation (KD) methods for link pr
ediction, i.e., predicted logit-based matching and node representation-based mat
ching. Upon observing direct KD analogs do not perform well for link prediction,
 we propose a relational KD framework, Linkless Link Prediction (LLP), to distil
l knowledge for link prediction with MLPs. Unlike simple KD methods that match i
ndependent link logits or node representations, LLP distills relational knowledg
e that is centered around each (anchor) node to the student MLP. Specifically, w
e propose rank-based matching and distribution-based matching strategies that co
mplement each other. Extensive experiments demonstrate that LLP boosts the link
prediction performance of MLPs with significant margins and even outperforms the

teacher GNNs on 7 out of 8 benchmarks. LLP also achieves a 70.68x speedup in link prediction inference compared to GNNs on the large-scale OGB dataset.

****

FedBR: Improving Federated Learning on Heterogeneous Data via Local Learning Bias Reduction

Yongxin Guo, Xiaoying Tang, Tao Lin

Federated Learning (FL) is a way for machines to learn from data that is kept locally, in order to protect the privacy of clients. This is typically done using local SGD, which helps to improve communication efficiency. However, such a scheme is currently constrained by slow and unstable convergence due to the variety of data on different clients' devices. In this work, we identify three under-explored phenomena of biased local learning that may explain these challenges caused by local updates in supervised FL. As a remedy, we propose FedBR, a novel unified algorithm that reduces the local learning bias on features and classifiers to tackle these challenges. FedBR has two components. The first component helps to reduce bias in local classifiers by balancing the output of the models. The second component helps to learn local features that are similar to global features, but different from those learned from other data sources. We conducted several experiments to test FedBR and found that it consistently outperforms other SOTA FL methods. Both of its components also individually show performance gains. Our code is available at https://github.com/lins-lab/fedbr.

****

Hierarchical Grammar-Induced Geometry for Data-Efficient Molecular Property Prediction

Minghao Guo, Veronika Thost, Samuel W Song, Adithya Balachandran, Payel Das, Jie Chen, Wojciech Matusik

The prediction of molecular properties is a crucial task in the field of material and drug discovery. The potential benefits of using deep learning techniques are reflected in the wealth of recent literature. Still, these techniques are faced with a common challenge in practice: Labeled data are limited by the cost of manual extraction from literature and laborious experimentation. In this work, we propose a data-efficient property predictor by utilizing a learnable hierarchical molecular grammar that can generate molecules from grammar production rules. Such a grammar induces an explicit geometry of the space of molecular graphs, which provides an informative prior on molecular structural similarity. The property prediction is performed using graph neural diffusion over the grammar-induced geometry. On both small and large datasets, our evaluation shows that this approach outperforms a wide spectrum of baselines, including supervised and pre-trained graph neural networks. We include a detailed ablation study and further analysis of our solution, showing its effectiveness in cases with extremely limited data.

****

Graph Neural Networks with Learnable and Optimal Polynomial Bases

Yuhe Guo, Zhewei Wei

Polynomial filters, a kind of Graph Neural Networks, typically use a predetermined polynomial basis and learn the coefficients from the training data. It has been observed that the effectiveness of the model is highly dependent on the property of the polynomial basis. Consequently, two natural and fundamental questions arise: Can we learn a suitable polynomial basis from the training data? Can we determine the optimal polynomial basis for a given graph and node features? In this paper, we propose two spectral GNN models that provide positive answers to the questions posed above. First, inspired by Favard's Theorem, we propose the FavardGNN model, which learns a polynomial basis from the space of all possible orthonormal bases. Second, we examine the supposedly unsolvable definition of optimal polynomial basis from Wang et al. (2022) and propose a simple model, OptBasisGNN, which computes the optimal basis for a given graph structure and graph signal. Extensive experiments are conducted to demonstrate the effectiveness of our proposed models. Our code is available at https://github.com/yuziGuo/FarOptBasis.

****

LongCoder: A Long-Range Pre-trained Language Model for Code Completion
Daya Guo, Canwen Xu, Nan Duan, Jian Yin, Julian Mcauley
In this paper, we introduce a new task for code completion that focuses on handling long code input and propose a sparse Transformer model, called LongCoder, to address this task. LongCoder employs a sliding window mechanism for self-attention and introduces two types of globally accessible tokens - bridge tokens and memory tokens - to improve performance and efficiency. Bridge tokens are inserted throughout the input sequence to aggregate local information and facilitate global interaction, while memory tokens are included to highlight important statements that may be invoked later and need to be memorized, such as package imports and definitions of classes, functions, or structures. We conduct experiments on a newly constructed dataset that contains longer code context and the publicly available CodeXGLUE benchmark. Experimental results demonstrate that LongCoder achieves superior performance on code completion tasks compared to previous models while maintaining comparable efficiency in terms of computational resources during inference.
****
Estimating Heterogeneous Treatment Effects: Mutual Information Bounds and Learning Algorithms
Xingzhuo Guo, Yuchen Zhang, Jianmin Wang, Mingsheng Long
Estimating heterogeneous treatment effects (HTE) from observational studies is rising in importance due to the widespread accumulation of data in many fields. Due to the selection bias behind the inaccessibility of counterfactual data, the problem differs fundamentally from supervised learning in a challenging way. However, existing works on modeling selection bias and corresponding algorithms do not naturally generalize to non-binary treatment spaces. To address this limitation, we propose to use mutual information to describe selection bias in estimating HTE and derive a novel error bound using the mutual information between the covariates and the treatments, which is the first error bound to cover general treatment schemes including multinoulli or continuous spaces. We then bring forth theoretically justified algorithms, the Mutual Information Treatment Network (MitNet), using adversarial optimization to reduce selection bias and obtain more accurate HTE estimations. Our algorithm reaches remarkable performance in both simulation study and empirical evaluation.
****
Identifying Useful Learnwares for Heterogeneous Label Spaces
Lan-Zhe Guo, Zhi Zhou, Yu-Feng Li, Zhi-Hua Zhou
The learnware paradigm aims to build a learnware market containing numerous learnwares, each of which is a well-performing machine learning model with a corresponding specification to describe its functionality so that future users can identify useful models for reuse according to their own requirements. With the learnware paradigm, model developers can spontaneously submit models to the market without leaking data privacy, and users can leverage models in the market to accomplish different machine learning tasks without having to build models from scratch. Recent studies have attempted to realize the model specification through Reduced Kernel Mean Embedding (RKME). In this paper, we make an attempt to improve the effectiveness of RKME specification for heterogeneous label spaces, where the learnware market does not contain a model that has the same label space as the user’s task, by considering a class-specific model specification explicitly, along with a class-wise learnware identification method. Both theoretical and empirical analyses show that our proposal can quickly and accurately find useful learnwares that satisfy users’ requirements. Moreover, we find that for a specific task, reusing a small model identified via the specification performs better than directly reusing a pre-trained generic big model.
****
High-dimensional Location Estimation via Norm Concentration for Subgamma Vectors
Shivam Gupta, Jasper C.H. Lee, Eric Price
In location estimation, we are given n samples from a known distribution f shifted by an unknown translation λ, and want to estimate λ as precisely as possible. Asymptotically, the maximum likelihood estimate achieves th

e Cramér-Rao bound of error $\mathcal N(0, \frac{1}{n\mathcal I})$, where $\mathcal I$ is the Fisher information of f. However, the n required for convergence depends on f, and may be arbitrarily large. We build on the theory using smoothed estimators to bound the error for finite n in terms of $\mathcal I_r$, the Fisher information of the r-smoothed distribution. As $n \to \infty$, $r \to 0$ at an explicit rate and this converges to the Cramér-Rao bound. We (1) improve the prior work for 1-dimensional f to converge for constant failure probability in addition to high probability, and (2) extend the theory to high-dimensional distributions. In the process, we prove a new bound on the norm of a high-dimensional random variable whose 1-dimensional projections are subgamma, which may be of independent interest.

****

GRAFENNE: Learning on Graphs with Heterogeneous and Dynamic Feature Sets

Shubham Gupta, Sahil Manchanda, Sayan Ranu, Srikanta J. Bedathur

Graph neural networks (GNNs), in general, are built on the assumption of a static set of features characterizing each node in a graph. This assumption is often violated in practice. Existing methods partly address this issue through feature imputation. However, these techniques (i) assume uniformity of feature set across nodes, (ii) are transductive by nature, and (iii) fail to work when features are added or removed over time. In this work, we address these limitations through a novel GNN framework called GRAFENNE. GRAFENNE performs a novel allotropic transformation on the original graph, wherein the nodes and features are decoupled through a bipartite encoding. Through a carefully chosen message passing framework on the allotropic transformation, we make the model parameter size independent of the number of features and thereby inductive to both unseen nodes and features. We prove that GRAFENNE is at least as expressive as any of the existing message-passing GNNs in terms of Weisfeiler-Leman tests, and therefore, the additional inductivity to unseen features does not come at the cost of expressivity. In addition, as demonstrated over four real-world graphs, GRAFENNE empowers the underlying GNN with high empirical efficacy and the ability to learn in continual fashion over streaming feature sets.

****

Online Platt Scaling with Calibeating

Chirag Gupta, Aaditya Ramdas

We present an online post-hoc calibration method, called Online Platt Scaling (OPS), which combines the Platt scaling technique with online logistic regression. We demonstrate that OPS smoothly adapts between i.i.d. and non-i.i.d. settings with distribution drift. Further, in scenarios where the best Platt scaling model is itself miscalibrated, we enhance OPS by incorporating a recently developed technique called calibeating to make it more robust. Theoretically, our resulting OPS+calibeating method is guaranteed to be calibrated for adversarial outcome sequences. Empirically, it is effective on a range of synthetic and real-world datasets, with and without distribution drifts, achieving superior performance without hyperparameter tuning. Finally, we extend all OPS ideas to the beta scaling method.

****

Multi-Task Structural Learning using Local Task Similarity induced Neuron Creation and Removal

Naresh Kumar Gurulingan, Bahram Zonooz, Elahe Arani

Multi-task learning has the potential to improve generalization by maximizing positive transfer between tasks while reducing task interference. Fully achieving this potential is hindered by manually designed architectures that remain static throughout training. On the contrary, learning in the brain occurs through structural changes that are in tandem with changes in synaptic strength. Thus, we propose Multi-Task Structural Learning (MTSL) that simultaneously learns the multi-task architecture and its parameters. MTSL begins with an identical single-task network for each task and alternates between a task-learning phase and a structural-learning phase. In the task learning phase, each network specializes in the corresponding task. In each of the structural learning phases, starting from the earliest layer, locally similar task layers first transfer their knowledge to

a newly created group layer before being removed. MTSL then uses the group layer in place of the corresponding removed task layers and moves on to the next layers. Our empirical results show that MTSL achieves competitive generalization with various baselines and improves robustness to out-of-distribution data.
****
Conditionally Strongly Log-Concave Generative Models
Florentin Guth, Etienne Lempereur, Joan Bruna, Stéphane Mallat
There is a growing gap between the impressive results of deep image generative models and classical algorithms that offer theoretical guarantees. The former suffer from mode collapse or memorization issues, limiting their application to scientific data. The latter require restrictive assumptions such as log-concavity to escape the curse of dimensionality. We partially bridge this gap by introducing conditionally strongly log-concave (CSLC) models, which factorize the data distribution into a product of conditional probability distributions that are strongly log-concave. This factorization is obtained with orthogonal projectors adapted to the data distribution. It leads to efficient parameter estimation and sampling algorithms, with theoretical guarantees, although the data distribution is not globally log-concave. We show that several challenging multiscale processes are conditionally log-concave using wavelet packet orthogonal projectors. Numerical results are shown for physical fields such as the φ^4 model and weak lensing convergence maps with higher resolution than in previous works.
****
DRew: Dynamically Rewired Message Passing with Delay
Benjamin Gutteridge, Xiaowen Dong, Michael M. Bronstein, Francesco Di Giovanni
Message passing neural networks (MPNNs) have been shown to suffer from the phenomenon of over-squashing that causes poor performance for tasks relying on long-range interactions. This can be largely attributed to message passing only occurring locally, over a node’s immediate neighbours. Rewiring approaches attempting to make graphs ’more connected’, and supposedly better suited to long-range tasks, often lose the inductive bias provided by distance on the graph since they make distant nodes communicate instantly at every layer. In this paper we propose a framework, applicable to any MPNN architecture, that performs a layer-dependent rewiring to ensure gradual densification of the graph. We also propose a delay mechanism that permits skip connections between nodes depending on the layer and their mutual distance. We validate our approach on several long-range tasks and show that it outperforms graph Transformers and multi-hop MPNNs.
****
Kernel Logistic Regression Approximation of an Understandable ReLU Neural Network
Marie Guyomard, Susana Barbosa, Lionel Fillatre
This paper proposes an understandable neural network whose score function is modeled as an additive sum of univariate spline functions. It extends usual understandable models like generative additive models, spline-based models, and neural additive models. It is shown that this neural network can be approximated by a logistic regression whose inputs are obtained with a non-linear preprocessing of input data. This preprocessing depends on the neural network initialization but this paper establishes that it can be replaced by a non random kernel-based preprocessing that no longer depends on the initialization. Hence, the convergence of the training process is guaranteed and the solution is unique for a given training dataset.
****
Conformal Prediction Sets for Graph Neural Networks
Soroush H. Zargarbashi, Simone Antonelli, Aleksandar Bojchevski
Despite the widespread use of graph neural networks (GNNs) we lack methods to reliably quantify their uncertainty. We propose a conformal procedure to equip GNNs with prediction sets that come with distribution-free guarantees – the output set contains the true label with arbitrarily high probability. Our post-processing procedure can wrap around any (pretrained) GNN, and unlike existing methods, results in meaningful sets even when the model provides only the top class. The key idea is to diffuse the node-wise conformity scores to incorporate neighborho

od information. By leveraging the network homophily we construct sets with compa
rable or better efficiency (average size) and significantly improved singleton h
it ratio (correct sets of size one). In addition to an extensive empirical evalu
ation, we investigate the theoretical conditions under which smoothing provably
improves efficiency.
****
Social learning spontaneously emerges by searching optimal heuristics with deep
reinforcement learning
Seungwoong Ha, Hawoong Jeong
How have individuals of social animals in nature evolved to learn from each othe
r, and what would be the optimal strategy for such learning in a specific enviro
nment? Here, we address both problems by employing a deep reinforcement learning
 model to optimize the social learning strategies (SLSs) of agents in a cooperat
ive game in a multi-dimensional landscape. Throughout the training for maximizin
g the overall payoff, we find that the agent spontaneously learns various concep
ts of social learning, such as copying, focusing on frequent and well-performing
 neighbors, self-comparison, long-term cooperation between agents, and the impor
tance of balancing between individual and social learning, without any explicit
guidance or prior knowledge about the system. The SLS from a fully trained agent
 outperforms all of the traditional, baseline SLSs in terms of mean payoff. We d
emonstrate the superior performance of the reinforcement learning agent in vario
us environments, including temporally changing environments and real social netw
orks, which also verifies the adaptability of our framework to different social
settings.
****
Convex Geometry of ReLU-layers, Injectivity on the Ball and Local Reconstruction
Daniel Haider, Martin Ehler, Peter Balazs
The paper uses a frame-theoretic setting to study the injectivity of a ReLU-laye
r on the closed ball of $\mathbb{R}^n$ and its non-negative part. In particular,
 the interplay between the radius of the ball and the bias vector is emphasized.
 Together with a perspective from convex geometry, this leads to a computational
ly feasible method of verifying the injectivity of a ReLU-layer under reasonable
 restrictions in terms of an upper bound of the bias vector. Explicit reconstruc
tion formulas are provided, inspired by the duality concept from frame theory. A
ll this gives rise to the possibility of quantifying the invertibility of a ReLU
-layer and a concrete reconstruction algorithm for any input vector on the ball.
****
Robust Counterfactual Explanations for Neural Networks With Probabilistic Guaran
tees
Faisal Hamman, Erfaun Noorani, Saumitra Mishra, Daniele Magazzeni, Sanghamitra D
utta
There is an emerging interest in generating robust counterfactual explanations t
hat would remain valid if the model is updated or changed even slightly. Towards
 finding robust counterfactuals, existing literature often assumes that the orig
inal model m and the new model M are bounded in the parameter space, i.e., $
\|\text{Params}(M){-}\text{Params}(m)\|{<}\Delta$. However, models can often cha
nge significantly in the parameter space with little to no change in their predi
ctions or accuracy on the given dataset. In this work, we introduce a mathematic
al abstraction termed naturally-occurring model change, which allows for arbitra
ry changes in the parameter space such that the change in predictions on points
that lie on the data manifold is limited. Next, we propose a measure – that we c
all Stability – to quantify the robustness of counterfactuals to potential model
 changes for differentiable models, e.g., neural networks. Our main contribution
 is to show that counterfactuals with sufficiently high value of Stability as de
fined by our measure will remain valid after potential “naturally-occurring” mod
el changes with high probability (leveraging concentration bounds for Lipschitz
function of independent Gaussians). Since our quantification depends on the loca
l Lipschitz constant around a data point which is not always available, we also
examine practical relaxations of our proposed measure and demonstrate experiment
ally how they can be incorporated to find robust counterfactuals for neural netw

orks that are close, realistic, and remain valid after potential model changes.
****
Wrapped Cauchy Distributed Angular Softmax for Long-Tailed Visual Recognition
Boran Han
Addressing imbalanced or long-tailed data is a major challenge in visual recogni tion tasks due to disparities between training and testing distributions and iss ues with data noise. We propose the Wrapped Cauchy Distributed Angular Softmax (WCDAS), a novel softmax function that incorporates data-wise Gaussian-based kern els into the angular correlation between feature representations and classifier weights, effectively mitigating noise and sparse sampling concerns. The class-wi se distribution of angular representation becomes a sum of these kernels. Our th eoretical analysis reveals that the wrapped Cauchy distribution excels the Gauss ian distribution in approximating mixed distributions. Additionally, WCDAS uses trainable concentration parameters to dynamically adjust the compactness and mar gin of each class. Empirical results confirm label-aware behavior in these param eters and demonstrate WCDAS's superiority over other state-of-the-art softmax-ba sed methods in handling long-tailed visual recognition across multiple benchmark datasets. The code is public available.
****
On the Impact of Knowledge Distillation for Model Interpretability
Hyeongrok Han, Siwon Kim, Hyun-Soo Choi, Sungroh Yoon
Several recent studies have elucidated why knowledge distillation (KD) improves model performance. However, few have researched the other advantages of KD in ad dition to its improving model performance. In this study, we have attempted to s how that KD enhances the interpretability as well as the accuracy of models. We measured the number of concept detectors identified in network dissection for a quantitative comparison of model interpretability. We attributed the improvement in interpretability to the class-similarity information transferred from the te acher to student models. First, we confirmed the transfer of class-similarity in formation from the teacher to student model via logit distillation. Then, we ana lyzed how class-similarity information affects model interpretability in terms o f its presence or absence and degree of similarity information. We conducted var ious quantitative and qualitative experiments and examined the results on differ ent datasets, different KD methods, and according to different measures of inter pretability. Our research showed that KD models by large models could be used mo re reliably in various fields. The code is available at https://github.com/Rok07 /KD_XAI.git.
****
Alternately Optimized Graph Neural Networks
Haoyu Han, Xiaorui Liu, Haitao Mao, Mohamadali Torkamani, Feng Shi, Victor Lee, Jiliang Tang
Graph Neural Networks (GNNs) have greatly advanced the semi-supervised node clas sification task on graphs. The majority of existing GNNs are trained in an end-t o-end manner that can be viewed as tackling a bi-level optimization problem. Thi s process is often inefficient in computation and memory usage. In this work, we propose a new optimization framework for semi-supervised learning on graphs fro m a multi-view learning perspective. The proposed framework can be conveniently solved by the alternating optimization algorithms, resulting in significantly im proved efficiency. Extensive experiments demonstrate that the proposed method ca n achieve comparable or better performance with state-of-the-art baselines while it has significantly better computation and memory efficiency.
****
System Identification of Neural Systems: If We Got It Right, Would We Know?
Yena Han, Tomaso A Poggio, Brian Cheung
Artificial neural networks are being proposed as models of parts of the brain. T he networks are compared to recordings of biological neurons, and good performan ce in reproducing neural responses is considered to support the model's validity . A key question is how much this system identification approach tells us about brain computation. Does it validate one model architecture over another? We eval uate the most commonly used comparison techniques, such as a linear encoding mod

el and centered kernel alignment, to correctly identify a model by replacing bra in recordings with known ground truth models. System identification performance is quite variable; it also depends significantly on factors independent of the g round truth architecture, such as stimuli images. In addition, we show the limit ations of using functional similarity scores in identifying higher-level archite ctural motifs.
****

Total Variation Graph Neural Networks
Jonas Berg Hansen, Filippo Maria Bianchi
Recently proposed Graph Neural Networks (GNNs) for vertex clustering are trained with an unsupervised minimum cut objective, approximated by a Spectral Clusteri ng (SC) relaxation. However, the SC relaxation is loose and, while it offers a c losed-form solution, it also yields overly smooth cluster assignments that poorl y separate the vertices. In this paper, we propose a GNN model that computes clu ster assignments by optimizing a tighter relaxation of the minimum cut based on graph total variation (GTV). The cluster assignments can be used directly to per form vertex clustering or to implement graph pooling in a graph classification f ramework. Our model consists of two core components: i) a message-passing layer that minimizes the ℓ_1 distance in the features of adjacent vertices, which is key to achieving sharp transitions between clusters; ii) an unsupervised los s function that minimizes the GTV of the cluster assignments while ensuring bala nced partitions. Experimental results show that our model outperforms other GNNs for vertex clustering and graph classification.
****

Learning Physical Models that Can Respect Conservation Laws
Derek Hansen, Danielle C. Maddix, Shima Alizadeh, Gaurav Gupta, Michael W. Mahon ey
Recent work in scientific machine learning (SciML) has focused on incorporating partial differential equation (PDE) information into the learning process. Much of this work has focused on relatively "easy” PDE operators (e.g., elliptic and parabolic), with less emphasis on relatively “hard” PDE operators (e.g., hyperbo lic). Within numerical PDEs, the latter problem class requires control of a type of volume element or conservation constraint, which is known to be challenging. Delivering on the promise of SciML requires seamlessly incorporating both types of problems into the learning process. To address this issue, we propose ProbCo nserv, a framework for incorporating constraints into a generic SciML architectu re. To do so, ProbConserv combines the integral form of a conservation law with a Bayesian update. We provide a detailed analysis of ProbConserv on learning wit h the Generalized Porous Medium Equation (GPME), a widely-applicable parameteriz ed family of PDEs that illustrates the qualitative properties of both easier and harder PDEs. ProbConserv is effective for easy GPME variants, performing well w ith state-of-the-art competitors; and for harder GPME variants it outperforms ot her approaches that do not guarantee volume conservation. ProbConserv seamlessly enforces physical conservation constraints, maintains probabilistic uncertainty quantification (UQ), and deals well with shocks and heteroscedasticity. In each case, it achieves superior predictive performance on downstream tasks.
****

On Pre-Training for Visuo-Motor Control: Revisiting a Learning-from-Scratch Base line
Nicklas Hansen, Zhecheng Yuan, Yanjie Ze, Tongzhou Mu, Aravind Rajeswaran, Hao S u, Huazhe Xu, Xiaolong Wang
In this paper, we examine the effectiveness of pre-training for visuo-motor cont rol tasks. We revisit a simple Learning-from-Scratch (LfS) baseline that incorpo rates data augmentation and a shallow ConvNet, and find that this baseline is su rprisingly competitive with recent approaches (PVR, MVP, R3M) that leverage froz en visual representations trained on large-scale vision datasets – across a vari ety of algorithms, task domains, and metrics in simulation and on a real robot. Our results demonstrate that these methods are hindered by a significant domain gap between the pre-training datasets and current benchmarks for visuo-motor con trol, which is alleviated by finetuning. Based on our findings, we provide recom

mendations for future research in pre-training for control and hope that our simple yet strong baseline will aid in accurately benchmarking progress in this area. Code: https://github.com/gemcollector/learning-from-scratch.

****

Leveraging Demonstrations to Improve Online Learning: Quality Matters

Botao Hao, Rahul Jain, Tor Lattimore, Benjamin Van Roy, Zheng Wen

We investigate the extent to which offline demonstration data can improve online learning. It is natural to expect some improvement, but the question is how, and by how much? We show that the degree of improvement must depend on the quality of the demonstration data. To generate portable insights, we focus on Thompson sampling (TS) applied to a multi-armed bandit as a prototypical online learning algorithm and model. The demonstration data is generated by an expert with a given competence level, a notion we introduce. We propose an informed TS algorithm that utilizes the demonstration data in a coherent way through Bayes' rule and derive a prior-dependent Bayesian regret bound. This offers insight into how pretraining can greatly improve online performance and how the degree of improvement increases with the expert's competence level. We also develop a practical, approximate informed TS algorithm through Bayesian bootstrapping and show substantial empirical regret reduction through experiments.

****

Coupled Variational Autoencoder

Xiaoran Hao, Patrick Shafto

Variational auto-encoders are powerful probabilistic models in generative tasks but suffer from generating low-quality samples which are caused by the holes in the prior. We propose the Coupled Variational Auto-Encoder (C-VAE), which formulates the VAE problem as one of Optimal Transport (OT) between the prior and data distributions. The C-VAE allows greater flexibility in priors and natural resolution of the prior hole problem by enforcing coupling between the prior and the data distribution and enables flexible optimization through the primal, dual, and semi-dual formulations of entropic OT. Simulations on synthetic and real data show that the C-VAE outperforms alternatives including VAE, WAE, and InfoVAE in fidelity to the data, quality of the latent representation, and in quality of generated samples.

****

GNOT: A General Neural Operator Transformer for Operator Learning

Zhongkai Hao, Zhengyi Wang, Hang Su, Chengyang Ying, Yinpeng Dong, Songming Liu, Ze Cheng, Jian Song, Jun Zhu

Learning partial differential equations' (PDEs) solution operators is an essential problem in machine learning. However, there are several challenges for learning operators in practical applications like the irregular mesh, multiple input functions, and complexity of the PDEs' solution. To address these challenges, we propose a general neural operator transformer (GNOT), a scalable and effective transformer-based framework for learning operators. By designing a novel heterogeneous normalized attention layer, our model is highly flexible to handle multiple input functions and irregular meshes. Besides, we introduce a geometric gating mechanism which could be viewed as a soft domain decomposition to solve the multi-scale problems. The large model capacity of the transformer architecture grants our model the possibility to scale to large datasets and practical problems. We conduct extensive experiments on multiple challenging datasets from different domains and achieve a remarkable improvement compared with alternative methods. Our code and data are publicly available at https://github.com/thu-ml/GNOT.

****

Algorithmic Collective Action in Machine Learning

Moritz Hardt, Eric Mazumdar, Celestine Mendler-Dünner, Tijana Zrnic

We initiate a principled study of algorithmic collective action on digital platforms that deploy machine learning algorithms. We propose a simple theoretical model of a collective interacting with a firm's learning algorithm. The collective pools the data of participating individuals and executes an algorithmic strategy by instructing participants how to modify their own data to achieve a collective goal. We investigate the consequences of this model in three fundamental lear

ning-theoretic settings: nonparametric optimal learning, parametric risk minimiz ation, and gradient-based optimization. In each setting, we come up with coordin ated algorithmic strategies and characterize natural success criteria as a funct ion of the collective’s size. Complementing our theory, we conduct systematic ex periments on a skill classification task involving tens of thousands of resumes from a gig platform for freelancers. Through more than two thousand model traini ng runs of a BERT-like language model, we see a striking correspondence emerge b etween our empirical observations and the predictions made by our theory. Taken together, our theory and experiments broadly support the conclusion that algorit hmic collectives of exceedingly small fractional size can exert significant cont rol over a platform’s learning algorithm.
****

Gaussian Process Priors for Systems of Linear Partial Differential Equations wit h Constant Coefficients

Marc Harkonen, Markus Lange-Hegermann, Bogdan Raita

Partial differential equations (PDEs) are important tools to model physical syst ems and including them into machine learning models is an important way of incor porating physical knowledge. Given any system of linear PDEs with constant coeff icients, we propose a family of Gaussian process (GP) priors, which we call EPGP , such that all realizations are exact solutions of this system. We apply the Eh renpreis-Palamodov fundamental principle, which works as a non-linear Fourier tr ansform, to construct GP kernels mirroring standard spectral methods for GPs. Ou r approach can infer probable solutions of linear PDE systems from any data such as noisy measurements, or pointwise defined initial and boundary conditions. Co nstructing EPGP-priors is algorithmic, generally applicable, and comes with a sp arse version (S-EPGP) that learns the relevant spectral frequencies and works be tter for big data sets. We demonstrate our approach on three families of systems of PDEs, the heat equation, wave equation, and Maxwell’s equations, where we im prove upon the state of the art in computation time and precision, in some exper iments by several orders of magnitude.
****

Theoretical Guarantees of Learning Ensembling Strategies with Applications to Ti me Series Forecasting

Hilaf Hasson, Danielle C. Maddix, Bernie Wang, Gaurav Gupta, Youngsuk Park

Ensembling is among the most popular tools in machine learning (ML) due to its e ffectiveness in minimizing variance and thus improving generalization. Most ense mbling methods for black-box base learners fall under the umbrella of "stacked g eneralization," namely training an ML algorithm that takes the inferences from t he base learners as input. While stacking has been widely applied in practice, i ts theoretical properties are poorly understood. In this paper, we prove a novel result, showing that choosing the best stacked generalization from a (finite or finite-dimensional) family of stacked generalizations based on cross-validated performance does not perform "much worse" than the oracle best. Our result stren gthens and significantly extends the results in Van der Laan et al. (2007). Insp ired by the theoretical analysis, we further propose a particular family of stac ked generalizations in the context of probabilistic forecasting, each one with a different sensitivity for how much the ensemble weights are allowed to vary acr oss items, timestamps in the forecast horizon, and quantiles. Experimental resul ts demonstrate the performance gain of the proposed method.
****

Global Context Vision Transformers

Ali Hatamizadeh, Hongxu Yin, Greg Heinrich, Jan Kautz, Pavlo Molchanov

We propose global context vision transformer (GC ViT), a novel architecture that enhances parameter and compute utilization for computer vision. Our method leve rages global context self-attention modules, joint with standard local self-atte ntion, to effectively and efficiently model both long and short-range spatial in teractions, without the need for expensive operations such as computing attentio n masks or shifting local windows. In addition, we address the lack of the induc tive bias in ViTs, and propose to leverage a modified fused inverted residual bl ocks in our architecture. Our proposed GC ViT achieves state-of-the-art results

across image classification, object detection and semantic segmentation tasks. On ImageNet-1K dataset for classification, the variants of GC ViT with 51M, 90M and 201M parameters achieve 84.3%, 85.0% and 85.7% Top-1 accuracy, respectively, at 224 image resolution and without any pre-training, hence surpassing comparably-sized prior art such as CNN-based ConvNeXt and ViT-based MaxViT and Swin Transformer by a large margin. Pre-trained GC ViT backbones in downstream tasks of object detection, instance segmentation, and semantic segmentation using MS COCO and ADE20K datasets outperform prior work consistently. Specifically, GC ViT with a 4-scale DINO detection head achieves a box AP of 58.3 on MS COCO dataset.
****

Counterfactual Analysis in Dynamic Latent State Models
Martin B Haugh, Raghav Singal
We provide an optimization-based framework to perform counterfactual analysis in a dynamic model with hidden states. Our framework is grounded in the “abduction, action, and prediction” approach to answer counterfactual queries and handles two key challenges where (1) the states are hidden and (2) the model is dynamic. Recognizing the lack of knowledge on the underlying causal mechanism and the possibility of infinitely many such mechanisms, we optimize over this space and compute upper and lower bounds on the counterfactual quantity of interest. Our work brings together ideas from causality, state-space models, simulation, and optimization, and we apply it on a breast cancer case study. To the best of our knowledge, we are the first to compute lower and upper bounds on a counterfactual query in a dynamic latent-state model.
****

Sampling-based Nyström Approximation and Kernel Quadrature
Satoshi Hayakawa, Harald Oberhauser, Terry Lyons
We analyze the Nyström approximation of a positive definite kernel associated with a probability measure. We first prove an improved error bound for the conventional Nyström approximation with i.i.d. sampling and singular-value decomposition in the continuous regime; the proof techniques are borrowed from statistical learning theory. We further introduce a refined selection of subspaces in Nyström approximation with theoretical guarantees that is applicable to non-i.i.d. landmark points. Finally, we discuss their application to convex kernel quadrature and give novel theoretical guarantees as well as numerical observations.
****

Width and Depth Limits Commute in Residual Networks
Soufiane Hayou, Greg Yang
We show that taking the width and depth to infinity in a deep neural network with skip connections, when branches are scaled by $1/\sqrt{depth}$, result in the same covariance structure no matter how that limit is taken. This explains why the standard infinite-width-then-depth approach provides practical insights even for networks with depth of the same order as width. We also demonstrate that the pre-activations, in this case, have Gaussian distributions which has direct applications in Bayesian deep learning. We conduct extensive simulations that show an excellent match with our theoretical findings.
****

A Generalization of ViT/MLP-Mixer to Graphs
Xiaoxin He, Bryan Hooi, Thomas Laurent, Adam Perold, Yann Lecun, Xavier Bresson
Graph Neural Networks (GNNs) have shown great potential in the field of graph representation learning. Standard GNNs define a local message-passing mechanism which propagates information over the whole graph domain by stacking multiple layers. This paradigm suffers from two major limitations, over-squashing and poor long-range dependencies, that can be solved using global attention but significantly increases the computational cost to quadratic complexity. In this work, we propose an alternative approach to overcome these structural limitations by leveraging the ViT/MLP-Mixer architectures introduced in computer vision. We introduce a new class of GNNs, called Graph ViT/MLP-Mixer, that holds three key properties. First, they capture long-range dependency and mitigate the issue of over-squashing as demonstrated on Long Range Graph Benchmark and TreeNeighbourMatch datasets. Second, they offer better speed and memory efficiency with a complexity lin

ear to the number of nodes and edges, surpassing the related Graph Transformer a nd expressive GNN models. Third, they show high expressivity in terms of graph i somorphism as they can distinguish at least 3-WL non-isomorphic graphs. We test our architecture on 4 simulated datasets and 7 real-world benchmarks, and show h ighly competitive results on all of them. The source code is available for repro ducibility at: https://github.com/XiaoxinHe/Graph-ViT-MLPMixer.
****
Domain Adaptation for Time Series Under Feature and Label Shifts
Huan He, Owen Queen, Teddy Koker, Consuelo Cuevas, Theodoros Tsiligkaridis, Mari nka Zitnik
Unsupervised domain adaptation (UDA) enables the transfer of models trained on s ource domains to unlabeled target domains. However, transferring complex time se ries models presents challenges due to the dynamic temporal structure variations across domains. This leads to feature shifts in the time and frequency represen tations. Additionally, the label distributions of tasks in the source and target domains can differ significantly, posing difficulties in addressing label shift s and recognizing labels unique to the target domain. Effectively transferring c omplex time series models remains a formidable problem. We present RAINCOAT, the first model for both closed-set and universal domain adaptation on complex time series. RAINCOAT addresses feature and label shifts by considering both tempora l and frequency features, aligning them across domains, and correcting for misal ignments to facilitate the detection of private labels. Additionally, RAINCOAT i mproves transferability by identifying label shifts in target domains. Our exper iments with 5 datasets and 13 state-of-the-art UDA methods demonstrate that RAIN COAT can improve transfer learning performance by up to 16.33% and can handle bo th closed-set and universal domain adaptation.
****
Contrastive Learning Meets Homophily: Two Birds with One Stone
Dongxiao He, Jitao Zhao, Rui Guo, Zhiyong Feng, Di Jin, Yuxiao Huang, Zhen Wang, Weixiong Zhang
Graph Contrastive Learning (GCL) has recently enjoyed great success as an effici ent self-supervised representation learning approach. However, the existing meth ods have focused on designing of contrastive modes and used data augmentation wi th a rigid and inefficient one-to-one sampling strategy. We adopted node neighbo rhoods to extend positive samplings and made avoided resorting to data augmentat ion to create different views. We also considered the homophily problem in Graph Neural Networks (GNNs) between the inter-class node pairs. The key novelty of o ur method hinged upon analyzing this GNNs problem and integrating the GCL sampli ng strategy with homophily discrimination, where we solved these two significant problems using one approach. We introduced a new parameterized neighbor samplin g component to replace the conventional sub-optimal samplings. By keeping and up dating the neighbor sets, both the positive sampling of GCL and the message pass ing of GNNs can be optimized. Moreover, we theoretically proved that the new met hod provided a lower bound of mutual information for unsupervised semantic learn ing, and it can also keep the lower bound with downstream tasks. In essence, our method is a new self-supervised approach, which we refer to as group discrimina tion, and it can make the downstream fine-tuning efficient. Our extensive empiri cal results demonstrate that the new method can significantly outperform the exi sting GCL methods because the former can solve the homophily problem in a self-s upervised way with the new group discrimination method used.
****
Nearly Minimax Optimal Reinforcement Learning for Linear Markov Decision Process es
Jiafan He, Heyang Zhao, Dongruo Zhou, Quanquan Gu
We study reinforcement learning (RL) with linear function approximation. For epi sodic time-inhomogeneous linear Markov decision processes (linear MDPs) whose tr ansition probability can be parameterized as a linear function of a given featur e mapping, we propose the first computationally efficient algorithm that achieve s the nearly minimax optimal regret $\tilde O(d\sqrt{H^3K})$, where d is the d imension of the feature mapping, H is the planning horizon, and K is the num

ber of episodes. Our algorithm is based on a weighted linear regression scheme w ith a carefully designed weight, which depends on a new variance estimator that (1) directly estimates the variance of the optimal value function, (2) monotonic ally decreases with respect to the number of episodes to ensure a better estimat ion accuracy, and (3) uses a rare-switching policy to update the value function estimator to control the complexity of the estimated value function class. Our w ork provides a complete answer to optimal RL with linear MDPs, and the developed algorithm and theoretical tools may be of independent interest.
****
CRISP: Curriculum based Sequential neural decoders for Polar code family
S Ashwin Hebbar, Viraj Vivek Nadkarni, Ashok Vardhan Makkuva, Suma Bhat, Sewoong Oh, Pramod Viswanath
Polar codes are widely used state-of-the-art codes for reliable communication th at have recently been included in the 5^{th} generation wireless standa rds (5G). However, there still remains room for design of polar decoders that are both efficient and reliable in the short blocklength regime. Motivated by re cent successes of data-driven channel decoders, we introduce a novel $\textbf{ C }$ur${\textbf{RI}}$culum based $\textbf{S}$equential neural decoder for $\textbf {P}$olar codes (CRISP). We design a principled curriculum, guided by information -theoretic insights, to train CRISP and show that it outperforms the successive- cancellation (SC) decoder and attains near-optimal reliability performance on th e $\text{Polar}(32,16)$ and $\text{Polar}(64,22)$ codes. The choice of the propo sed curriculum is critical in achieving the accuracy gains of CRISP, as we show by comparing against other curricula. More notably, CRISP can be readily extende d to Polarization-Adjusted-Convolutional (PAC) codes, where existing SC decoders are significantly less reliable. To the best of our knowledge, CRISP constructs the first data-driven decoder for PAC codes and attains near-optimal performanc e on the $\text{PAC}(32,16)$ code.
****
Sketch-Flip-Merge: Mergeable Sketches for Private Distinct Counting
Jonathan Hehir, Daniel Ting, Graham Cormode
Data sketching is a critical tool for distinct counting, enabling multisets to b e represented by compact summaries that admit fast cardinality estimates. Becaus e sketches may be merged to summarize multiset unions, they are a basic building block in data warehouses. Although many practical sketches for cardinality esti mation exist, none provide privacy when merging. We propose the first practical cardinality sketches that are simultaneously mergeable, differentially private (DP), and have low empirical errors. These introduce a novel randomized algorithm for performing logical operations on noisy bits, a tight privacy analysis, and provably optimal estimation. Our sketches dramatically outperform existing theor etical solutions in simulations and on real-world data.
****
Functional Neural Networks: Shift invariant models for functional data with appl ications to EEG classification
Florian Heinrichs, Mavin Heim, Corinna Weber
It is desirable for statistical models to detect signals of interest independent ly of their position. If the data is generated by some smooth process, this addi tional structure should be taken into account. We introduce a new class of neura l networks that are shift invariant and preserve smoothness of the data: functio nal neural networks (FNNs). For this, we use methods from functional data analys is (FDA) to extend multi-layer perceptrons and convolutional neural networks to functional data. We propose different model architectures, show that the models outperform a benchmark model from FDA in terms of accuracy and successfully use FNNs to classify electroencephalography (EEG) data.
****
Distance Weighted Supervised Learning for Offline Interaction Data
Joey Hejna, Jensen Gao, Dorsa Sadigh
Sequential decision making algorithms often struggle to leverage different sourc es of unstructured offline interaction data. Imitation learning (IL) methods bas ed on supervised learning are robust, but require optimal demonstrations, which

are hard to collect. Offline goal-conditioned reinforcement learning (RL) algori thms promise to learn from sub-optimal data, but face optimization challenges es pecially with high-dimensional data. To bridge the gap between IL and RL, we int roduce Distance Weighted Supervised Learning or DWSL, a supervised method for le arning goal-conditioned policies from offline data. DWSL models the entire distr ibution of time-steps between states in offline data with only supervised learni ng, and uses this distribution to approximate shortest path distances. To extrac t a policy, we weight actions by their reduction in distance estimates. Theoreti cally, DWSL converges to an optimal policy constrained to the data distribution, an attractive property for offline learning, without any bootstrapping. Across all datasets we test, DWSL empirically maintains behavior cloning as a lower bou nd while still exhibiting policy improvement. In high-dimensional image domains, DWSL surpasses the performance of both prior goal-conditioned IL and RL algorit hms. Visualizations and code can be found at https://sites.google.com/view/dwsl/ home.

****

Group Equivariant Fourier Neural Operators for Partial Differential Equations
Jacob Helwig, Xuan Zhang, Cong Fu, Jerry Kurtin, Stephan Wojtowytsch, Shuiwang J i

We consider solving partial differential equations (PDEs) with Fourier neural op erators (FNOs), which operate in the frequency domain. Since the laws of physics do not depend on the coordinate system used to describe them, it is desirable t o encode such symmetries in the neural operator architecture for better performa nce and easier learning. While encoding symmetries in the physical domain using group theory has been studied extensively, how to capture symmetries in the freq uency domain is under-explored. In this work, we extend group convolutions to th e frequency domain and design Fourier layers that are equivariant to rotations, translations, and reflections by leveraging the equivariance property of the Fou rier transform. The resulting G-FNO architecture generalizes well across input resolutions and performs well in settings with varying levels of symmetry. Our code is publicly available as part of the AIRS library (https://github.com/divel ab/AIRS).

****

Training-Free Neural Active Learning with Initialization-Robustness Guarantees
Apivich Hemachandra, Zhongxiang Dai, Jasraj Singh, See-Kiong Ng, Bryan Kian Hsia ng Low

Existing neural active learning algorithms have aimed to optimize the predictive performance of neural networks (NNs) by selecting data for labelling. However, other than a good predictive performance, being robust against random parameter initializations is also a crucial requirement in safety-critical applications. T o this end, we introduce our expected variance with Gaussian processes (EV-GP) c riterion for neural active learning, which is theoretically guaranteed to select data points which lead to trained NNs with both (a) good predictive performance s and (b) initialization robustness. Importantly, our EV-GP criterion is trainin g-free, i.e., it does not require any training of the NN during data selection, which makes it computationally efficient. We empirically demonstrate that our EV -GP criterion is highly correlated with both initialization robustness and gener alization performance, and show that it consistently outperforms baseline method s in terms of both desiderata, especially in situations with limited initial dat a or large batch sizes.

****

A Study of Global and Episodic Bonuses for Exploration in Contextual MDPs
Mikael Henaff, Minqi Jiang, Roberta Raileanu

Exploration in environments which differ across episodes has received increasing attention in recent years. Current methods use some combination of global novel ty bonuses, computed using the agent’s entire training experience, and episodic novelty bonuses, computed using only experience from the current episode. Howeve r, the use of these two types of bonuses has been ad-hoc and poorly understood. In this work, we shed light on the behavior of these two types of bonuses throug h controlled experiments on easily interpretable tasks as well as challenging pi

xel-based settings. We find that the two types of bonuses succeed in different s ettings, with episodic bonuses being most effective when there is little shared structure across episodes and global bonuses being effective when more structure is shared. We develop a conceptual framework which makes this notion of shared structure precise by considering the variance of the value function across conte xts, and which provides a unifying explanation of our empirical results. We furt hermore find that combining the two bonuses can lead to more robust performance across different degrees of shared structure, and investigate different algorith mic choices for defining and combining global and episodic bonuses based on func tion approximation. This results in an algorithm which sets a new state of the a rt across 16 tasks from the MiniHack suite used in prior work, and also performs robustly on Habitat and Montezuma’s Revenge.

****

Robust Camera Pose Refinement for Multi-Resolution Hash Encoding

Hwan Heo, Taekyung Kim, Jiyoung Lee, Jaewon Lee, Soohyun Kim, Hyunwoo J. Kim, Ji n-Hwa Kim

Multi-resolution hash encoding has recently been proposed to reduce the computat ional cost of neural renderings, such as NeRF. This method requires accurate cam era poses for the neural renderings of given scenes. However, contrary to previo us methods jointly optimizing camera poses and 3D scenes, the naive gradient-bas ed camera pose refinement method using multi-resolution hash encoding severely d eteriorates performance. We propose a joint optimization algorithm to calibrate the camera pose and learn a geometric representation using efficient multi-resol ution hash encoding. Showing that the oscillating gradient flows of hash encodin g interfere with the registration of camera poses, our method addresses the issu e by utilizing smooth interpolation weighting to stabilize the gradient oscillat ion for the ray samplings across hash grids. Moreover, the curriculum training p rocedure helps to learn the level-wise hash encoding, further increasing the pos e refinement. Experiments on the novel-view synthesis datasets validate that our learning frameworks achieve state-of-the-art performance and rapid convergence of neural rendering.

****

Generalized Teacher Forcing for Learning Chaotic Dynamics

Florian Hess, Zahra Monfared, Manuel Brenner, Daniel Durstewitz

Chaotic dynamical systems (DS) are ubiquitous in nature and society. Often we ar e interested in reconstructing such systems from observed time series for predic tion or mechanistic insight, where by reconstruction we mean learning geometrica l and invariant temporal properties of the system in question (like attractors). However, training reconstruction algorithms like recurrent neural networks (RNN s) on such systems by gradient-descent based techniques faces severe challenges. This is mainly due to exploding gradients caused by the exponential divergence of trajectories in chaotic systems. Moreover, for (scientific) interpretability we wish to have as low dimensional reconstructions as possible, preferably in a model which is mathematically tractable. Here we report that a surprisingly simp le modification of teacher forcing leads to provably strictly all-time bounded g radients in training on chaotic systems, and, when paired with a simple architec tural rearrangement of a tractable RNN design, piecewise-linear RNNs (PLRNNs), a llows for faithful reconstruction in spaces of at most the dimensionality of the observed system. We show on several DS that with these amendments we can recons truct DS better than current SOTA algorithms, in much lower dimensions. Performa nce differences were particularly compelling on real world data with which most other methods severely struggled. This work thus led to a simple yet powerful DS reconstruction algorithm which is highly interpretable at the same time.

****

Causal Modeling of Policy Interventions From Treatment-Outcome Sequences

Ça■lar H■zl■, S. T. John, Anne Tuulikki Juuti, Tuure Tapani Saarinen, Kirsi Hann ele Pietiläinen, Pekka Marttinen

A treatment policy defines when and what treatments are applied to affect some o utcome of interest. Data-driven decision-making requires the ability to predict what happens if a policy is changed. Existing methods that predict how the outco

me evolves under different scenarios assume that the tentative sequences of futu re treatments are fixed in advance, while in practice the treatments are determi ned stochastically by a policy and may depend, for example, on the efficiency of previous treatments. Therefore, the current methods are not applicable if the t reatment policy is unknown or a counterfactual analysis is needed. To handle the se limitations, we model the treatments and outcomes jointly in continuous time, by combining Gaussian processes and point processes. Our model enables the esti mation of a treatment policy from observational sequences of treatments and outc omes, and it can predict the interventional and counterfactual progression of th e outcome after an intervention on the treatment policy (in contrast with the ca usal effect of a single treatment). We show with real-world and semi-synthetic d ata on blood glucose progression that our method can answer causal queries more accurately than existing alternatives.

****

Monotonicity and Double Descent in Uncertainty Estimation with Gaussian Processe s

Liam Hodgkinson, Chris Van Der Heide, Fred Roosta, Michael W. Mahoney

Despite their importance for assessing reliability of predictions, uncertainty q uantification (UQ) measures in machine learning models have only recently begun to be rigorously characterized. One prominent issue is the curse of dimensionali ty: it is commonly believed that the marginal likelihood should be reminiscent o f cross-validation metrics and both should deteriorate with larger input dimensi ons. However, we prove that by tuning hyperparameters to maximize marginal likel ihood (the empirical Bayes procedure), performance, as measured by the marginal likelihood, improves monotonically with the input dimension. On the other hand, cross-validation metrics exhibit qualitatively different behavior that is charac teristic of double descent. Cold posteriors, which have recently attracted inter est due to their improved performance in certain settings, appear to exacerbate these phenomena. We verify empirically that our results hold for real data, beyo nd our considered assumptions, and we explore consequences involving synthetic c ovariates.

****

AdaBoost is not an Optimal Weak to Strong Learner

Mikael Møller Høgsgaard, Kasper Green Larsen, Martin Ritzert

AdaBoost is a classic boosting algorithm for combining multiple inaccurate class ifiers produced by a weak learner, to produce a strong learner with arbitrarily high accuracy when given enough training data. Determining the optimal number of samples necessary to obtain a given accuracy of the strong learner, is a basic learning theoretic question. Larsen and Ritzert (NeurIPS’22) recently presented the first provably optimal weak-to-strong learner. However, their algorithm is s omewhat complicated and it remains an intriguing question whether the prototypic al boosting algorithm AdaBoost also makes optimal use of training samples. In th is work, we answer this question in the negative. Concretely, we show that the s ample complexity of AdaBoost, and other classic variations thereof, are sub-opti mal by at least one logarithmic factor in the desired accuracy of the strong lea rner.

****

Dual Propagation: Accelerating Contrastive Hebbian Learning with Dyadic Neurons

Rasmus Høier, D. Staudt, Christopher Zach

Activity difference based learning algorithms—such as contrastive Hebbian learni ng and equilibrium propagation—have been proposed as biologically plausible alte rnatives to error back-propagation. However, on traditional digital chips these algorithms suffer from having to solve a costly inference problem twice, making these approaches more than two orders of magnitude slower than back-propagation. In the analog realm equilibrium propagation may be promising for fast and energ y efficient learning, but states still need to be inferred and stored twice. Ins pired by lifted neural networks and compartmental neuron models we propose a sim ple energy based compartmental neuron model, termed dual propagation, in which e ach neuron is a dyad with two intrinsic states. At inference time these intrinsi c states encode the error/activity duality through their difference and their me

an respectively. The advantage of this method is that only a single inference ph
ase is needed and that inference can be solved in layerwise closed-form. Experim
entally we show on common computer vision datasets, including Imagenet32x32, tha
t dual propagation performs equivalently to back-propagation both in terms of ac
curacy and runtime.
****
Multi-Task Off-Policy Learning from Bandit Feedback
Joey Hong, Branislav Kveton, Manzil Zaheer, Sumeet Katariya, Mohammad Ghavamzade
h
Many practical problems involve solving similar tasks. In recommender systems, t
he tasks can be users with similar preferences; in search engines, the tasks can
 be items with similar affinities. To learn statistically efficiently, the tasks
 can be organized in a hierarchy, where the task affinity is captured using an u
nknown latent parameter. We study the problem of off-policy learning for similar
 tasks from logged bandit feedback. To solve the problem, we propose a hierarchi
cal off-policy optimization algorithm HierOPO. The key idea is to estimate the t
ask parameters using the hierarchy and then act pessimistically with respect to
them. To analyze the algorithm, we develop novel Bayesian error bounds. Our boun
ds are the first in off-policy learning that improve with a more informative pri
or and capture statistical gains due to hierarchical models. Therefore, they are
 of a general interest. HierOPO also performs well in practice. Our experiments
demonstrate the benefits of using the hierarchy over solving each task independe
ntly.
****
Constrained Optimization via Exact Augmented Lagrangian and Randomized Iterative
 Sketching
Ilgee Hong, Sen Na, Michael W. Mahoney, Mladen Kolar
We consider solving equality-constrained nonlinear, nonconvex optimization probl
ems. This class of problems appears widely in a variety of applications in machi
ne learning and engineering, ranging from constrained deep neural networks, to o
ptimal control, to PDE-constrained optimization. We develop an adaptive inexact
Newton method for this problem class. In each iteration, we solve the Lagrangian
 Newton system inexactly via a randomized iterative sketching solver, and select
 a suitable stepsize by performing line search on an exact augmented Lagrangian
merit function. The randomized solvers have advantages over deterministic linear
 system solvers by significantly reducing per-iteration flops complexity and sto
rage cost, when equipped with suitable sketching matrices. Our method adaptively
 controls the accuracy of the randomized solver and the penalty parameters of th
e exact augmented Lagrangian, to ensure that the inexact Newton direction is a d
escent direction of the exact augmented Lagrangian. This allows us to establish
a global almost sure convergence. We also show that a unit stepsize is admissibl
e locally, so that our method exhibits a local linear convergence. Furthermore,
we prove that the linear convergence can be strengthened to superlinear converge
nce if we gradually sharpen the adaptive accuracy condition on the randomized so
lver. We demonstrate the superior performance of our method on benchmark nonline
ar problems in CUTEst test set, constrained logistic regression with data from L
IBSVM, and a PDE-constrained problem.
****
Revisiting Data-Free Knowledge Distillation with Poisoned Teachers
Junyuan Hong, Yi Zeng, Shuyang Yu, Lingjuan Lyu, Ruoxi Jia, Jiayu Zhou
Data-free knowledge distillation (KD) helps transfer knowledge from a pre-traine
d model (known as the teacher model) to a smaller model (known as the student mo
del) without access to the original training data used for training the teacher
model. However, the security of the synthetic or out-of-distribution (OOD) data
required in data-free KD is largely unknown and under-explored. In this work, we
 make the first effort to uncover the security risk of data-free KD w.r.t. untru
sted pre-trained models. We then propose Anti-Backdoor Data-Free KD (ABD), the f
irst plug-in defensive method for data-free KD methods to mitigate the chance of
 potential backdoors being transferred. We empirically evaluate the effectivenes
s of our proposed ABD in diminishing transferred backdoor knowledge while mainta

ining compatible downstream performances as the vanilla KD. We envision this wor
k as a milestone for alarming and mitigating the potential backdoors in data-fre
e KD. Codes are released at https://github.com/illidanlab/ABD .
****

simple diffusion: End-to-end diffusion for high resolution images
Emiel Hoogeboom, Jonathan Heek, Tim Salimans
Currently, applying diffusion models in pixel space of high resolution images is
 difficult. Instead, existing approaches focus on diffusion in lower dimensional
 spaces (latent diffusion), or have multiple super-resolution levels of generati
on referred to as cascades. The downside is that these approaches add additional
 complexity to the diffusion framework. This paper aims to improve denoising dif
fusion for high resolution images while keeping the model as simple as possible.
 The paper is centered around the research question: How can one train a standar
d denoising diffusion models on high resolution images, and still obtain perform
ance comparable to these alternate approaches? The four main findings are: 1) th
e noise schedule should be adjusted for high resolution images, 2) It is suffici
ent to scale only a particular part of the architecture, 3) dropout should be ad
ded at specific locations in the architecture, and 4) downsampling is an effecti
ve strategy to avoid high resolution feature maps. Combining these simple yet ef
fective techniques, we achieve state-of-the-art on image generation among diffus
ion models without sampling modifiers on ImageNet.
****

Causal Strategic Classification: A Tale of Two Shifts
Guy Horowitz, Nir Rosenfeld
When users can benefit from certain predictive outcomes, they may be prone to ac
t to achieve those outcome, e.g., by strategically modifying their features. The
 goal in strategic classification is therefore to train predictive models that a
re robust to such behavior. However, the conventional framework assumes that cha
nging features does not change actual outcomes, which depicts users as "gaming"
the system. Here we remove this assumption, and study learning in a causal strat
egic setting where true outcomes do change. Focusing on accuracy as our primary
objective, we show how strategic behavior and causal effects underlie two comple
menting forms of distribution shift. We characterize these shifts, and propose a
 learning algorithm that balances between these two forces and over time, and pe
rmits end-to-end training. Experiments on synthetic and semi-synthetic data demo
nstrate the utility of our approach.
****

Fair and Accurate Decision Making through Group-Aware Learning
Ramtin Hosseini, Li Zhang, Bhanu Garg, Pengtao Xie
The integration of machine learning models in various real-world applications is
 becoming more prevalent to assist humans in their daily decision-making tasks a
s a result of recent advancements in this field. However, it has been discovered
 that there is a tradeoff between the accuracy and fairness of these decision-ma
king tasks. In some cases, these AI systems can be unfair by exhibiting bias or
discrimination against certain social groups, which can have severe consequences
 in real life. Inspired by one of the most well-known human learning skills call
ed grouping, we address this issue by proposing a novel machine learning (ML) fr
amework where the ML model learns to group a diverse set of problems into distin
ct subgroups to solve each subgroup using its specific sub-model. Our proposed f
ramework involves three stages of learning, which are formulated as a three-leve
l optimization problem: 1) grouping problems into subgroups, 2) learning group-s
pecific sub-models for problem-solving, and 3) updating group assignments of tra
ining examples by minimizing validation loss. These three learning stages are pe
rformed end-to-end in a joint manner using gradient descent. To improve fairness
 and accuracy, we develop an efficient optimization algorithm to solve this thre
e-level optimization problem. To further decrease the risk of overfitting in sma
ll datasets using our LBG method, we incorporate domain adaptation techniques in
 the second stage of training. We further apply our method to differentiable neu
ral architecture search (NAS) methods.
****

Approximation Algorithms for Fair Range Clustering
Sedjro Salomon Hotegni, Sepideh Mahabadi, Ali Vakilian
This paper studies the fair range clustering problem in which the data points ar e from different demographic groups and the goal is to pick k centers with the minimum clustering cost such that each group is at least minimally represented in the centers set and no group dominates the centers set. More precisely, given a set of n points in a metric space (P, d) where each point belongs to one of the ℓ different demographics (i.e., $P = P_1 \uplus P_2 \uplus \cdots \u plus P_\ell$) and a set of ℓ intervals $[\alpha_1, \beta_1], \cdots, [\alph a_\ell, \beta_\ell]$ on desired number of centers from each group, the goal is t o pick a set of k centers C with minimum ℓ_p-clustering cost (i.e., $(\ sum_{v\in P} d(v,C)^p)^{1/p}$) such that for each group $i\in \ell$, $|C\cap P_i | \in [\alpha_i, \beta_i]$. In particular, the fair range ℓ_p-clustering ca ptures fair range k-center, k-median and k-means as its special cases. In this work, we provide an efficient constant factor approximation algorithm for t he fair range ℓ_p-clustering for all values of $p\in [1,\infty)$.
****
Decoding Layer Saliency in Language Transformers
Elizabeth Mary Hou, Gregory David Castanon
In this paper, we introduce a strategy for identifying textual saliency in large -scale language models applied to classification tasks. In visual networks where saliency is more well-studied, saliency is naturally localized through the conv olutional layers of the network; however, the same is not true in modern transfo rmer-stack networks used to process natural language. We adapt gradient-based sa liency methods for these networks, propose a method for evaluating the degree of semantic coherence of each layer, and demonstrate consistent improvement over n umerous other methods for textual saliency on multiple benchmark classification datasets. Our approach requires no additional training or access to labelled dat a, and is comparatively very computationally efficient.
****
PromptBoosting: Black-Box Text Classification with Ten Forward Passes
Bairu Hou, Joe O’Connor, Jacob Andreas, Shiyu Chang, Yang Zhang
We describe PromptBoosting, a query-efficient procedure for building a text clas sifier from a neural language model (LM) without access to the LM’s parameters, gradients, or hidden representations. This form of "black-box" classifier traini ng has become increasingly important as the cost of training and inference in la rge-scale LMs has grown. But existing black-box LM classifier learning approache s are themselves computationally inefficient, typically specializing LMs to the target task by searching in a large space of (discrete or continuous) prompts us ing zeroth-order optimization methods. Instead of directly optimizing in prompt space, PromptBoosting obtains a small pool of prompts via a gradient-free approa ch and then constructs a large pool of weak learners by pairing these prompts wi th different elements of the LM’s output distribution. These weak learners are t hen ensembled using the AdaBoost algorithm. The entire learning process requires only a small number of forward passes and no backward pass. Experiments show th at PromptBoosting achieves state-of-the-art performance in multiple black-box fe w-shot classification tasks, and matches or outperforms full fine-tuning in both few-shot and standard learning paradigms, while training 10x faster than existi ng black-box methods.
****
Sparse Learning of Dynamical Systems in RKHS: An Operator-Theoretic Approach
Boya Hou, Sina Sanjari, Nathan Dahlin, Subhonmesh Bose, Umesh Vaidya
Transfer operators provide a rich framework for representing the dynamics of ver y general, nonlinear dynamical systems. When interacting with reproducing kernel Hilbert spaces (RKHS), descriptions of dynamics often incur prohibitive data st orage requirements, motivating dataset sparsification as a precursory step to co mputation. Further, in practice, data is available in the form of trajectories, introducing correlation between samples. In this work, we present a method for s parse learning of transfer operators from β-mixing stochastic processes, i n both discrete and continuous time, and provide sample complexity analysis exte

nding existing theoretical guarantees for learning from non-sparse, i.i.d. data. In addressing continuous-time settings, we develop precise descriptions using covariance-type operators for the infinitesimal generator that aids in the sample complexity analysis. We empirically illustrate the efficacy of our sparse embedding approach through deterministic and stochastic nonlinear system examples.

****

Probably Anytime-Safe Stochastic Combinatorial Semi-Bandits

Yunlong Hou, Vincent Y. F. Tan, Zixin Zhong

Motivated by concerns about making online decisions that incur undue amount of risk at each time step, in this paper, we formulate the probably anytime-safe stochastic combinatorial semi-bandits problem. In this problem, the agent is given the option to select a subset of size at most K from a set of L ground items. Each item is associated to a certain mean reward as well as a variance that represents its risk. To mitigate the risk that the agent incurs, we require that with probability at least $1-\delta$, over the entire horizon of time T, each of the choices that the agent makes should contain items whose sum of variances does not exceed a certain variance budget. We call this probably anytime-safe constraint. Under this constraint, we design and analyze an algorithm PASCombUCB that minimizes the regret over the horizon of time T. By developing accompanying information-theoretic lower bounds, we show that under both the problem-dependent and problem-independent paradigms, PASCombUCB is almost asymptotically optimal. Experiments are conducted to corroborate our theoretical findings. Our problem setup, the proposed PASCombUCB algorithm, and novel analyses are applicable to domains such as recommendation systems and transportation in which an agent is allowed to choose multiple items at a single time step and wishes to control the risk over the whole time horizon.

****

Automatic Data Augmentation via Invariance-Constrained Learning

Ignacio Hounie, Luiz F. O. Chamon, Alejandro Ribeiro

Underlying data structures, such as symmetries or invariance to transformations, are often exploited to improve the solution of learning tasks. However, embedding these properties in models or learning algorithms can be challenging and computationally intensive. Data augmentation, on the other hand, induces these symmetries during training by applying multiple transformations to the input data. Despite its ubiquity, its effectiveness depends on the choices of which transformations to apply, when to do so, and how often. In fact, there is both empirical and theoretical evidence that the indiscriminate use of data augmentation can introduce biases that outweigh its benefits. This work tackles these issues by automatically adapting the data augmentation while solving the learning task. To do so, it formulates data augmentation as an invariance constrained learning problem and leverages Monte Carlo Markov Chain (MCMC) sampling to solve it. The result is an algorithm that not only does away with a priori searches for augmentation distributions, but also dynamically controls if and when data augmentation is applied. We validate empirically our theoretical developments in automatic data augmentation benchmarks for CIFAR and ImageNet-100 datasets. Furthermore, our experiments show how this approach can be used to gather insights on the actual symmetries underlying a learning task.

****

Thompson Sampling with Diffusion Generative Prior

Yu-Guan Hsieh, Shiva Kasiviswanathan, Branislav Kveton, Patrick Blöbaum

In this work, we initiate the idea of using denoising diffusion models to learn priors for online decision making problems. We specifically focus on bandit meta-learning, aiming to learn a policy that performs well across bandit tasks of a same class. To this end, we train a diffusion model that learns the underlying task distribution and combine Thompson sampling with the learned prior to deal with new tasks at test time. Our posterior sampling algorithm carefully balances between the learned prior and the noisy observations that come from the learner’s interaction with the environment. To capture realistic bandit scenarios, we propose a novel diffusion model training procedure that trains from incomplete and noisy data, which could be of independent interest. Finally, our extensive exper

iments clearly demonstrate the potential of the proposed approach.
****
Tighter Analysis for ProxSkip
Zhengmian Hu, Heng Huang
In this paper, we provide a tighter analysis for ProxSkip, an algorithm that allows fewer proximal operator computations to solve composite optimization problems. We improve the existing decreasing speed of Lyapunov function from $\mathcal{O}(p^2)$ to $\mathcal{O}(p)$, when p, the frequency of the proximal operators is small enough. Our theoretical analysis also reveals the drawbacks of using large step sizes for gradient descent in ProxSkip when the proximal operator part is the bottleneck. Our main motivation comes from the continuous limit in which the original analysis of ProxSkip fails to guarantee convergence when both the step size γ and frequency p tend to zero. We construct a counterexample to demonstrate why such counterintuitive behavior occurs for the original analysis and then propose a novel Lyapunov function variant to construct a tighter analysis, avoiding the problem of the old one. Such a new Lyapunov function can be directly extended to many other variants of ProxSkip. When applied to stochastic gradient setup, our analysis leads to an improved proximal operator complexity for SProxSkip from $\mathcal{O}(\sqrt{\frac{1}{\varepsilon\mu^2}}\log(\frac{1}{\varepsilon}))$ to $\mathcal{O}(\sqrt{\kappa}\log(\frac{1}{\varepsilon}))$.
****
Omnipredictors for Constrained Optimization
Lunjia Hu, Inbal Rachel Livni Navon, Omer Reingold, Chutong Yang
The notion of omnipredictors (Gopalan, Kalai, Reingold, Sharan and Wieder ITCS 2022), suggested a new paradigm for loss minimization. Rather than learning a predictor based on a known loss function, omnipredictors can easily be post-processed to minimize any one of a rich family of loss functions compared with the loss of hypotheses in a class $\mathcal C$. It has been shown that such omnipredictors exist and are implied (for all convex and Lipschitz loss functions) by the notion of multicalibration from the algorithmic fairness literature. In this paper, we introduce omnipredictors for constrained optimization and study their complexity and implications. The notion that we introduce allows the learner to be unaware of the loss function that will be later assigned as well as the constraints that will be later imposed, as long as the subpopulations that are used to define these constraints are known. We show how to obtain omnipredictors for constrained optimization problems, relying on appropriate variants of multicalibration. We also investigate the implications of this notion when the constraints used are so-called group fairness notions.
****
GFlowNet-EM for Learning Compositional Latent Variable Models
Edward J Hu, Nikolay Malkin, Moksh Jain, Katie E Everett, Alexandros Graikos, Yoshua Bengio
Latent variable models (LVMs) with discrete compositional latents are an important but challenging setting due to a combinatorially large number of possible configurations of the latents. A key tradeoff in modeling the posteriors over latents is between expressivity and tractable optimization. For algorithms based on expectation-maximization (EM), the E-step is often intractable without restrictive approximations to the posterior. We propose the use of GFlowNets, algorithms for sampling from an unnormalized density by learning a stochastic policy for sequential construction of samples, for this intractable E-step. By training GFlowNets to sample from the posterior over latents, we take advantage of their strengths as amortized variational inference algorithms for complex distributions over discrete structures. Our approach, GFlowNet-EM, enables the training of expressive LVMs with discrete compositional latents, as shown by experiments on non-context-free grammar induction and on images using discrete variational autoencoders (VAEs) without conditional independence enforced in the encoder.
****
Blockwise Stochastic Variance-Reduced Methods with Parallel Speedup for Multi-Block Bilevel Optimization
Quanqi Hu, Zi-Hao Qiu, Zhishuai Guo, Lijun Zhang, Tianbao Yang

In this paper, we consider non-convex multi-block bilevel optimization (MBBO) pr
oblems, which involve $m\gg 1$ lower level problems and have important applicati
ons in machine learning. Designing a stochastic gradient and controlling its var
iance is more intricate due to the hierarchical sampling of blocks and data and
the unique challenge of estimating hyper-gradient. We aim to achieve three nice
properties for our algorithm: (a) matching the state-of-the-art complexity of st
andard BO problems with a single block; (b) achieving parallel speedup by sampli
ng I blocks and sampling B samples for each sampled block per-iteration; (c)
 avoiding the computation of the inverse of a high-dimensional Hessian matrix es
timator. However, it is non-trivial to achieve all of these by observing that ex
isting works only achieve one or two of these properties. To address the involve
d challenges for achieving (a, b, c), we propose two stochastic algorithms by us
ing advanced blockwise variance-reduction techniques for tracking the Hessian ma
trices (for low-dimensional problems) or the Hessian-vector products (for high-d
imensional problems), and prove an iteration complexity of $O(\frac{m\epsilon^{-
3}\mathbb{I}(I \textless m)}{I\sqrt{I}}+\frac{m\epsilon^{-3}}{I\sqrt{B}})$ for f
inding an ϵ-stationary point under appropriate conditions. We also cond
uct experiments to verify the effectiveness of the proposed algorithms comparing
 with existing MBBO algorithms.

****

Language Instructed Reinforcement Learning for Human-AI Coordination

Hengyuan Hu, Dorsa Sadigh

One of the fundamental quests of AI is to produce agents that coordinate well wi
th humans. This problem is challenging, especially in domains that lack high qua
lity human behavioral data, because multi-agent reinforcement learning (RL) ofte
n converges to different equilibria from the ones that humans prefer. We propose
 a novel framework, instructRL, that enables humans to specify what kind of stra
tegies they expect from their AI partners through natural language instructions.
 We use pretrained large language models to generate a prior policy conditioned
on the human instruction and use the prior to regularize the RL objective. This
leads to the RL agent converging to equilibria that are aligned with human prefe
rences. We show that instructRL converges to human-like policies that satisfy th
e given instructions in a proof-of-concept environment as well as the challengin
g Hanabi benchmark. Finally, we show that knowing the language instruction signi
ficantly boosts human-AI coordination performance in human evaluations in Hanabi
.

****

Surface Snapping Optimization Layer for Single Image Object Shape Reconstruction

Yuan-Ting Hu, Alex Schwing, Raymond A. Yeh

Reconstructing the 3D shape of objects observed in a single image is a challengi
ng task. Recent approaches rely on visual cues extracted from a given image lear
ned from a deep net. In this work, we leverage recent advances in monocular scen
e understanding to incorporate an additional geometric cue of surface normals. F
or this, we proposed a novel optimization layer that encourages the face normals
 of the reconstructed shape to be aligned with estimated surface normals. We dev
elop a computationally efficient conjugate-gradient-based method that avoids the
 computation of a high-dimensional sparse matrix. We show this framework to achi
eve compelling shape reconstruction results on the challenging Pix3D and ShapeNe
t datasets.

****

Learning to Learn from APIs: Black-Box Data-Free Meta-Learning

Zixuan Hu, Li Shen, Zhenyi Wang, Baoyuan Wu, Chun Yuan, Dacheng Tao

Data-free meta-learning (DFML) aims to enable efficient learning of new tasks by
 meta-learning from a collection of pre-trained models without access to the tra
ining data. Existing DFML work can only meta-learn from (i) white-box and (ii) s
mall-scale pre-trained models (iii) with the same architecture, neglecting the m
ore practical setting where the users only have inference access to the APIs wit
h arbitrary model architectures and model scale inside. To solve this issue, we
propose a Bi-level Data-free Meta Knowledge Distillation (BiDf-MKD) framework to
 transfer more general meta knowledge from a collection of black-box APIs to one

single meta model. Specifically, by just querying APIs, we inverse each API to recover its training data via a zero-order gradient estimator and then perform meta-learning via a novel bi-level meta knowledge distillation structure, in which we design a boundary query set recovery technique to recover a more informative query set near the decision boundary. In addition, to encourage better generalization within the setting of limited API budgets, we propose task memory replay to diversify the underlying task distribution by covering more interpolated tasks. Extensive experiments in various real-world scenarios show the superior performance of our BiDf-MKD framework.

****

For Pre-Trained Vision Models in Motor Control, Not All Policy Learning Methods are Created Equal

Yingdong Hu, Renhao Wang, Li Erran Li, Yang Gao

In recent years, increasing attention has been directed to leveraging pre-trained vision models for motor control. While existing works mainly emphasize the importance of this pre-training phase, the arguably equally important role played by downstream policy learning during control-specific fine-tuning is often neglected. It thus remains unclear if pre-trained vision models are consistent in their effectiveness under different control policies. To bridge this gap in understanding, we conduct a comprehensive study on 14 pre-trained vision models using 3 distinct classes of policy learning methods, including reinforcement learning (RL), imitation learning through behavior cloning (BC), and imitation learning with a visual reward function (VRF). Our study yields a series of intriguing results, including the discovery that the effectiveness of pre-training is highly dependent on the choice of the downstream policy learning algorithm. We show that conventionally accepted evaluation based on RL methods is highly variable and therefore unreliable, and further advocate for using more robust methods like VRF and BC. To facilitate more universal evaluations of pre-trained models and their policy learning methods in the future, we also release a benchmark of 21 tasks across 3 different environments alongside our work.

****

Beyond Lipschitz Smoothness: A Tighter Analysis for Nonconvex Optimization

Zhengmian Hu, Xidong Wu, Heng Huang

Negative and positive curvatures affect optimization in different ways. However, a lot of existing optimization theories are based on the Lipschitz smoothness assumption, which cannot differentiate between the two. In this paper, we propose to use two separate assumptions for positive and negative curvatures, so that we can study the different implications of the two. We analyze the Lookahead and Local SGD methods as concrete examples. Both of them require multiple copies of model parameters and communication among them for every certain period of time in order to prevent divergence. We show that the minimum communication frequency is inversely proportional to the negative curvature, and when the negative curvature becomes zero, we recover the existing theory results for convex optimization. Finally, both experimentally and theoretically, we demonstrate that modern neural networks have highly unbalanced positive/negative curvatures. Thus, an analysis based on separate positive and negative curvatures is more pertinent.

****

Understanding the Impact of Adversarial Robustness on Accuracy Disparity

Yuzheng Hu, Fan Wu, Hongyang Zhang, Han Zhao

While it has long been empirically observed that adversarial robustness may be at odds with standard accuracy and may have further disparate impacts on different classes, it remains an open question to what extent such observations hold and how the class imbalance plays a role within. In this paper, we attempt to understand this question of accuracy disparity by taking a closer look at linear classifiers under a Gaussian mixture model. We decompose the impact of adversarial robustness into two parts: an inherent effect that will degrade the standard accuracy on all classes due to the robustness constraint, and the other caused by the class imbalance ratio, which will increase the accuracy disparity compared to standard training. Furthermore, we also show that such effects extend beyond the Gaussian mixture model, by generalizing our data model to the general family of

stable distributions. More specifically, we demonstrate that while the constrai
nt of adversarial robustness consistently degrades the standard accuracy in the
balanced class setting, the class imbalance ratio plays a fundamentally differen
t role in accuracy disparity compared to the Gaussian case, due to the heavy tai
l of the stable distribution. We additionally perform experiments on both synthe
tic and real-world datasets to corroborate our theoretical findings. Our empiric
al results also suggest that the implications may extend to nonlinear models ove
r real-world datasets. Our code is publicly available on GitHub at https://githu
b.com/Accuracy-Disparity/AT-on-AD.
****

Reinforcement Learning in Low-rank MDPs with Density Features
Audrey Huang, Jinglin Chen, Nan Jiang
MDPs with low-rank transitions—that is, the transition matrix can be factored in
to the product of two matrices, left and right—is a highly representative struct
ure that enables tractable learning. The left matrix enables expressive function
approximation for value-based learning and has been studied extensively. In thi
s work, we instead investigate sample-efficient learning with density features,
i.e., the right matrix, which induce powerful models for state-occupancy distrib
utions. This setting not only sheds light on leveraging unsupervised learning in
RL, but also enables plug-in solutions for settings like convex RL. In the offl
ine setting, we propose an algorithm for off-policy estimation of occupancies th
at can handle non-exploratory data. Using this as a subroutine, we further devis
e an online algorithm that constructs exploratory data distributions in a level-
by-level manner. As a central technical challenge, the additive error of occupan
cy estimation is incompatible with the multiplicative definition of data coverag
e. In the absence of strong assumptions like reachability, this incompatibility
easily leads to exponential error blow-up, which we overcome via novel technical
tools. Our results also readily extend to the representation learning setting,
when the density features are unknown and must be learned from an exponentially
large candidate set.
****

Composer: Creative and Controllable Image Synthesis with Composable Conditions
Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, Jingren Zhou
Recent large-scale generative models learned on big data are capable of synthesi
zing incredible images yet suffer from limited controllability. This work offers
a new generation paradigm that allows flexible control of the output image, suc
h as spatial layout and palette, while maintaining the synthesis quality and mod
el creativity. With compositionality as the core idea, we first decompose an ima
ge into representative factors, and then train a diffusion model with all these
factors as the conditions to recompose the input. At the inference stage, the ri
ch intermediate representations work as composable elements, leading to a huge d
esign space (i.e., exponentially proportional to the number of decomposed factor
s) for customizable content creation. It is noteworthy that our approach, which
we call Composer, supports various levels of conditions, such as text descriptio
n as the global information, depth map and sketch as the local guidance, color h
istogram for low-level details, etc. Besides improving controllability, we confi
rm that Composer serves as a general framework and facilitates a wide range of c
lassical generative tasks without retraining. Code and models will be made avail
able.
****

Model-Aware Contrastive Learning: Towards Escaping the Dilemmas
Zizheng Huang, Haoxing Chen, Ziqi Wen, Chao Zhang, Huaxiong Li, Bo Wang, Chunlin
Chen
Contrastive learning (CL) continuously achieves significant breakthroughs across
multiple domains. However, the most common InfoNCE-based methods suffer from so
me dilemmas, such as uniformity-tolerance dilemma (UTD) and gradient reduction,
both of which are related to a $\mathcal{P}_{ij}$ term. It has been identified t
hat UTD can lead to unexpected performance degradation. We argue that the fixity
of temperature is to blame for UTD. To tackle this challenge, we enrich the CL
loss family by presenting a Model-Aware Contrastive Learning (MACL) strategy, wh

ose temperature is adaptive to the magnitude of alignment that reflects the basi c confidence of the instance discrimination task, then enables CL loss to adjust the penalty strength for hard negatives adaptively. Regarding another dilemma, the gradient reduction issue, we derive the limits of an involved gradient scali ng factor, which allows us to explain from a unified perspective why some recent approaches are effective with fewer negative samples, and summarily present a g radient reweighting to escape this dilemma. Extensive remarkable empirical resul ts in vision, sentence, and graph modality validate our approach’s general impro vement for representation learning and downstream tasks.
****

High-dimensional Clustering onto Hamiltonian Cycle
Tianyi Huang, Shenghui Cheng, Stan Z. Li, Zhengjun Zhang
Clustering aims to group unlabelled samples based on their similarities and is w idespread in high-dimensional data analysis. However, most of the clustering met hods merely generate pseudo labels and thus are unable to simultaneously present the similarities between different clusters and outliers. This paper proposes a new framework called High-dimensional Clustering onto Hamiltonian Cycle (HCHC) to solve the above problems. First, HCHC combines global structure with local st ructure in one objective function for deep clustering, improving the labels as r elative probabilities, to mine the similarities between different clusters while keeping the local structure in each cluster. Then, the anchors of different clu sters are sorted on the optimal Hamiltonian cycle generated by the cluster simil arities and mapped on the circumference of a circle. Finally, a sample with a hi gher probability of a cluster will be mapped closer to the corresponding anchor. In this way, our framework allows us to appreciate three aspects visually and s imultaneously - clusters (formed by samples with high probabilities), cluster si milarities (represented as circular distances), and outliers (recognized as dots far away from all clusters). The theoretical analysis and experiments illustrat e the superiority of HCHC.
****

Banker Online Mirror Descent: A Universal Approach for Delayed Online Bandit Lea rning
Jiatai Huang, Yan Dai, Longbo Huang
We propose Banker Online Mirror Descent (Banker-OMD), a novel framework generali zing the classical Online Mirror Descent (OMD) technique in the online learning literature. The Banker-OMD framework almost completely decouples feedback delay handling and the task-specific OMD algorithm design, thus facilitating the desig n of new algorithms capable of efficiently and robustly handling feedback delays . Specifically, it offers a general methodology for achieving $\widetilde{\mathc al O}(\sqrt{T} + \sqrt{D})$-style regret bounds in online bandit learning tasks with delayed feedback, where T is the number of rounds and D is the total fe edback delay. We demonstrate the power of Banker-OMD by applications to two impo rtant bandit learning scenarios with delayed feedback, including delayed scale-f ree adversarial Multi-Armed Bandits (MAB) and delayed adversarial linear bandits . Banker-OMD leads to the first delayed scale-free adversarial MAB algorithm ach ieving $\widetilde{\mathcal O}(\sqrt{K}L(\sqrt T+\sqrt D))$ regret and the first delayed adversarial linear bandit algorithm achieving $\widetilde{\mathcal O}(\ text{poly}(n)(\sqrt{T} + \sqrt{D}))$ regret. As a corollary, the first applicati on also implies $\widetilde{\mathcal O}(\sqrt{KT}L)$ regret for non-delayed scal e-free adversarial MABs, which is the first to match the $\Omega(\sqrt{KT}L)$ lo wer bound up to logarithmic factors and can be of independent interest.
****

Fast Algorithms for Distributed k-Clustering with Outliers
Junyu Huang, Qilong Feng, Ziyun Huang, Jinhui Xu, Jianxin Wang
In this paper, we study the k-clustering problems with outliers in distributed setting. The current best results for the distributed k-center problem with o utliers have quadratic local running time with communication cost dependent on t he aspect ratio Δ of the given instance, which may constraint the scalabi lity of the algorithms for handling large-scale datasets. To achieve better comm unication cost for the problem with faster local running time, we propose an inl

iers-recalling sampling method, which avoids guessing the optimal radius of the given instance, and can achieve a 4-round bi-criteria $(14(1+\epsilon),1+\epsilon)$-approximation with linear local running time in the data size and communication cost independent of the aspect ratio. To obtain a more practical algorithm for the problem, we propose another space-narrowing sampling method, which automatically adjusts the sample size to adapt to different outliers distributions on each machine, and can achieve a 2-round bi-criteria $(14(1+\epsilon),1+\epsilon)$-approximation with communication cost independent of the number of outliers. We show that, if the data points are randomly partitioned across machines, our proposed sampling-based methods can be extended to the k-median/means problems with outliers, and can achieve $(O(\frac{1}{\epsilon^2}),1+\epsilon)$-approximation with communication cost independent of the number of outliers. Empirical experiments suggest that the proposed 2-round distributed algorithms outperform other state-of-the-art algorithms.

****

Searching Large Neighborhoods for Integer Linear Programs with Contrastive Learning

Taoan Huang, Aaron M Ferber, Yuandong Tian, Bistra Dilkina, Benoit Steiner

Integer Linear Programs (ILPs) are powerful tools for modeling and solving a large number of combinatorial optimization problems. Recently, it has been shown that Large Neighborhood Search (LNS), as a heuristic algorithm, can find high-quality solutions to ILPs faster than Branch and Bound. However, how to find the right heuristics to maximize the performance of LNS remains an open problem. In this paper, we propose a novel approach, CL-LNS, that delivers state-of-the-art anytime performance on several ILP benchmarks measured by metrics including the primal gap, the primal integral, survival rates and the best performing rate. Specifically, CL-LNS collects positive and negative solution samples from an expert heuristic that is slow to compute and learns a more efficient one with contrastive learning. We use graph attention networks and a richer set of features to further improve its performance.

****

On Coresets for Clustering in Small Dimensional Euclidean spaces

Lingxiao Huang, Ruiyuan Huang, Zengfeng Huang, Xuan Wu

We consider the problem of constructing small coresets for k-Median in Euclidean spaces. Given a large set of data points $P\subset \mathbb{R}^d$, a coreset is a much smaller set $S\subset \mathbb{R}^d$, so that the k-Median costs of any k centers w.r.t. P and S are close. Existing literature mainly focuses on the high-dimension case and there has been great success in obtaining dimension-independent bounds, whereas the case for small d is largely unexplored. Considering many applications of Euclidean clustering algorithms are in small dimensions and the lack of systematic studies in the current literature, this paper investigates coresets for k-Median in small dimensions. For small d, a natural question is whether existing near-optimal dimension-independent bounds can be significantly improved. We provide affirmative answers to this question for a range of parameters. Moreover, new lower bound results are also proved, which are the highest for small d. In particular, we completely settle the coreset size bound for 1-d k-Median (up to log factors). Interestingly, our results imply a strong separation between 1-d 1-Median and 1-d 2-Median. As far as we know, this is the first such separation between $k=1$ and $k=2$ in any dimension.

****

Make-An-Audio: Text-To-Audio Generation with Prompt-Enhanced Diffusion Models

Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, Zhou Zhao

Large-scale multimodal generative modeling has created milestones in text-to-image and text-to-video generation. Its application to audio still lags behind for two main reasons: the lack of large-scale datasets with high-quality text-audio pairs, and the complexity of modeling long continuous audio data. In this work, we propose Make-An-Audio with a prompt-enhanced diffusion model that addresses these gaps by 1) introducing pseudo prompt enhancement with a distill-then-reprog

ram approach, it alleviates data scarcity with orders of magnitude concept compo
sitions by using language-free audios; 2) leveraging spectrogram autoencoder to
predict the self-supervised audio representation instead of waveforms. Together
with robust contrastive language-audio pretraining (CLAP) representations, Make-
An-Audio achieves state-of-the-art results in both objective and subjective benc
hmark evaluation. Moreover, we present its controllability and generalization fo
r X-to-Audio with "No Modality Left Behind", for the first time unlocking the ab
ility to generate high-definition, high-fidelity audios given a user-defined mod
ality input. Audio samples are available at https://Make-An-Audio.github.io
****

The Power of Uniform Sampling for k-Median
Lingxiao Huang, Shaofeng H.-C. Jiang, Jianing Lou
We study the power of uniform sampling for k-Median in various metric spaces.
We relate the query complexity for approximating k-Median, to a key parameter
of the dataset, called the balancedness $\beta \in (0, 1]$ (with 1 being perfe
ctly balanced). We show that any algorithm must make $\Omega(1 / \beta)$ queries
 to the point set in order to achieve $O(1)$-approximation for k-Median. This
particularly implies existing constructions of coresets, a popular data reductio
n technique, cannot be query-efficient. On the other hand, we show a simple unif
orm sample of $\mathrm{poly}(k \epsilon^{-1} \beta^{-1})$ points suffices for $(
1 + \epsilon)$-approximation for k-Median for various metric spaces, which nea
rly matches the lower bound. We conduct experiments to verify that in many real
datasets, the balancedness parameter is usually well bounded, and that the unifo
rm sampling performs consistently well even for the case with moderately large b
alancedness, which justifies that uniform sampling is indeed a viable approach f
or solving k-Median.
****

Reparameterized Policy Learning for Multimodal Trajectory Optimization
Zhiao Huang, Litian Liang, Zhan Ling, Xuanlin Li, Chuang Gan, Hao Su
We investigate the challenge of parametrizing policies for reinforcement learnin
g (RL) in high-dimensional continuous action spaces. Our objective is to develop
 a multimodal policy that overcomes limitations inherent in the commonly-used Ga
ussian parameterization. To achieve this, we propose a principled framework that
 models the continuous RL policy as a generative model of optimal trajectories.
By conditioning the policy on a latent variable, we derive a novel variational b
ound as the optimization objective, which promotes exploration of the environmen
t. We then present a practical model-based RL method, called Reparameterized Pol
icy Gradient (RPG), which leverages the multimodal policy parameterization and l
earned world model to achieve strong exploration capabilities and high data effi
ciency. Empirical results demonstrate that our method can help agents evade loca
l optima in tasks with dense rewards and solve challenging sparse-reward environ
ments by incorporating an object-centric intrinsic reward. Our method consistent
ly outperforms previous approaches across a range of tasks. Code and supplementa
ry materials are available on the project page https://haosulab.github.io/RPG/
****

Theoretical Bounds on the Network Community Profile from Low-rank Semi-definite
Programming
Yufan Huang, C. Seshadhri, David F. Gleich
We study a new connection between a technical measure called μ-conductance t
hat arises in the study of Markov chains for sampling convex bodies and the netw
ork community profile that characterizes size-resolved properties of clusters an
d communities in social and information networks. The idea of μ-conductance
is similar to the traditional graph conductance, but disregards sets with small
volume. We derive a sequence of optimization problems including a low-rank semi-
definite program from which we can derive a lower bound on the optimal μ-con
ductance value. These ideas give the first theoretically sound bound on the beha
vior of the network community profile for a wide range of cluster sizes. The alg
orithm scales up to graphs with hundreds of thousands of nodes and we demonstrat
e how our framework validates the predicted structures of real-world graphs.
****

NeuralStagger: Accelerating Physics-constrained Neural PDE Solver with Spatial-temporal Decomposition
Xinquan Huang, Wenlei Shi, Qi Meng, Yue Wang, Xiaotian Gao, Jia Zhang, Tie-Yan Liu
Neural networks have shown great potential in accelerating the solution of partial differential equations (PDEs). Recently, there has been a growing interest in introducing physics constraints into training neural PDE solvers to reduce the use of costly data and improve the generalization ability. However, these physics constraints, based on certain finite dimensional approximations over the function space, must resolve the smallest scaled physics to ensure the accuracy and stability of the simulation, resulting in high computational costs from large input, output, and neural networks. This paper proposes a general acceleration methodology called NeuralStagger by spatially and temporally decomposing the original learning tasks into several coarser-resolution subtasks. We define a coarse-resolution neural solver for each subtask, which requires fewer computational resources, and jointly train them with the vanilla physics-constrained loss by simply arranging their outputs to reconstruct the original solution. Due to the perfect parallelism between them, the solution is achieved as fast as a coarse-resolution neural solver. In addition, the trained solvers bring the flexibility of simulating with multiple levels of resolution. We demonstrate the successful application of NeuralStagger on 2D and 3D fluid dynamics simulations, which leads to an additional $10\sim100\times$ speed-up. Moreover, the experiment also shows that the learned model could be well used for optimal control.
****
Policy Contrastive Imitation Learning
Jialei Huang, Zhao-Heng Yin, Yingdong Hu, Yang Gao
Adversarial imitation learning (AIL) is a popular method that has recently achieved much success. However, the performance of AIL is still unsatisfactory on the more challenging tasks. We find that one of the major reasons is due to the low quality of AIL discriminator representation. Since the AIL discriminator is trained via binary classification that does not necessarily discriminate the policy from the expert in a meaningful way, the resulting reward might not be meaningful either. We propose a new method called Policy Contrastive Imitation Learning (PCIL) to resolve this issue. PCIL learns a contrastive representation space by anchoring on different policies and uses a smooth cosine-similarity-based reward to encourage imitation learning. Our proposed representation learning objective can be viewed as a stronger version of the AIL objective and provide a more meaningful comparison between the agent and the policy. From a theoretical perspective, we show the validity of our method using the apprenticeship learning framework. Furthermore, our empirical evaluation on the DeepMind Control suite demonstrates that PCIL can achieve state-of-the-art performance. Finally, qualitative results suggest that PCIL builds a smoother and more meaningful representation space for imitation learning.
****
Are Large Kernels Better Teachers than Transformers for ConvNets?
Tianjin Huang, Lu Yin, Zhenyu Zhang, Li Shen, Meng Fang, Mykola Pechenizkiy, Zhangyang Wang, Shiwei Liu
This paper reveals a new appeal of the recently emerged large-kernel Convolutional Neural Networks (ConvNets): as the teacher in Knowledge Distillation (KD) for small-kernel ConvNets. While Transformers have led state-of-the-art (SOTA) performance in various fields with ever-larger models and labeled data, small-kernel ConvNets are considered more suitable for resource-limited applications due to the efficient convolution operation and compact weight sharing. KD is widely used to boost the performance of small-kernel ConvNets. However, previous research shows that it is not quite effective to distill knowledge (e.g., global information) from Transformers to small-kernel ConvNets, presumably due to their disparate architectures. We hereby carry out a first-of-its-kind study unveiling that modern large-kernel ConvNets, a compelling competitor to Vision Transformers, are remarkably more effective teachers for small-kernel ConvNets, due to more similar architectures. Our findings are backed up by extensive experiments on both lo

git-level and feature-level KD "out of the box", with no dedicated architectural nor training recipe modifications. Notably, we obtain the best-ever pure ConvNet under 30M parameters with 83.1% top-1 accuracy on ImageNet, outperforming current SOTA methods including ConvNeXt V2 and Swin V2. We also find that beneficial characteristics of large-kernel ConvNets, e.g., larger effective receptive fields, can be seamlessly transferred to students through this large-to-small kernel distillation. Code is available at: https://github.com/VITA-Group/SLaK.
****
Achieving Linear Speedup in Non-IID Federated Bilevel Learning
Minhui Huang, Dewei Zhang, Kaiyi Ji
Federated bilevel learning has received increasing attention in various emerging machine learning and communication applications. Recently, several Hessian-vector-based algorithms have been proposed to solve the federated bilevel optimization problem. However, several important properties in federated learning such as the partial client participation and the linear speedup for convergence (i.e., the convergence rate and complexity are improved linearly with respect to the number of sampled clients) in the presence of non-i.i.d. datasets, still remain open. In this paper, we fill these gaps by proposing a new federated bilevel algorithm named FedMBO with a novel client sampling scheme in the federated hypergradient estimation. We show that FedMBO achieves a convergence rate of $\mathcal{O}\big(\frac{1}{\sqrt{nK}}+\frac{1}{K}+\frac{\sqrt{n}}{K^{3/2}}\big)$ on non-i.i.d. datasets, where n is the number of participating clients in each round, and K is the total number of iteration. This is the first theoretical linear speedup result for non-i.i.d. federated bilevel optimization. Extensive experiments validate our theoretical results and demonstrate the effectiveness of our proposed method.
****
Federated Linear Contextual Bandits with User-level Differential Privacy
Ruiquan Huang, Huanyu Zhang, Luca Melis, Milan Shen, Meisam Hejazinia, Jing Yang
This paper studies federated linear contextual bandits under the notion of user-level differential privacy (DP). We first introduce a unified federated bandits framework that can accommodate various definitions of DP in the sequential decision-making setting. We then formally introduce user-level central DP (CDP) and local DP (LDP) in the federated bandits framework, and investigate the fundamental trade-offs between the learning regrets and the corresponding DP guarantees in a federated linear contextual bandits model. For CDP, we propose a federated algorithm termed as $\texttt{ROBIN}$ and show that it is near-optimal in terms of the number of clients M and the privacy budget ε by deriving nearly-matching upper and lower regret bounds when user-level DP is satisfied. For LDP, we obtain several lower bounds, indicating that learning under user-level (ε,δ)-LDP must suffer a regret blow-up factor at least $\min\{1/\varepsilon,M\}$ or $\min\{1/\sqrt{\varepsilon},\sqrt{M}\}$ under different conditions.
****
Straightening Out the Straight-Through Estimator: Overcoming Optimization Challenges in Vector Quantized Networks
Minyoung Huh, Brian Cheung, Pulkit Agrawal, Phillip Isola
This work examines the challenges of training neural networks using vector quantization using straight-through estimation. We find that the main cause of training instability is the discrepancy between the model embedding and the code-vector distribution. We identify the factors that contribute to this issue, including the codebook gradient sparsity and the asymmetric nature of the commitment loss, which leads to misaligned code-vector assignments. We propose to address this issue via affine re-parameterization of the code vectors. Additionally, we introduce an alternating optimization to reduce the gradient error introduced by the straight-through estimation. Moreover, we propose an improvement to the commitment loss to ensure better alignment between the codebook representation and the model embedding. These optimization methods improve the mathematical approximation of the straight-through estimation and, ultimately, the model performance. We demonstrate the effectiveness of our methods on several common model architectur

es, such as AlexNet, ResNet, and ViT, across various tasks, including image clas
sification and generative modeling.
****

Cut your Losses with Squentropy
Like Hui, Mikhail Belkin, Stephen Wright
Nearly all practical neural models for classification are trained using the cros
s-entropy loss. Yet this ubiquitous choice is supported by little theoretical or
 empirical evidence. Recent work (Hui & Belkin, 2020) suggests that training usi
ng the (rescaled) square loss is often superior in terms of the classification a
ccuracy. In this paper we propose the "squentropy" loss, which is the sum of two
 terms: the cross-entropy loss and the average square loss over the incorrect cl
asses. We provide an extensive set of experiment on multi-class classification p
roblems showing that the squentropy loss outperforms both the pure cross-entropy
 and rescaled square losses in terms of the classification accuracy. We also dem
onstrate that it provides significantly better model calibration than either of
these alternative losses and, furthermore, has less variance with respect to the
 random initialization. Additionally, in contrast to the square loss, squentropy
 loss can frequently be trained using exactly the same optimization parameters,
including the learning rate, as the standard cross-entropy loss, making it a tru
e ”plug-and-play” replacement. Finally, unlike the rescaled square loss, multicl
ass squentropy contains no parameters that need to be adjusted.
****

SOM-CPC: Unsupervised Contrastive Learning with Self-Organizing Maps for Structu
red Representations of High-Rate Time Series
Iris A.M. Huijben, Arthur Andreas Nijdam, Sebastiaan Overeem, Merel M Van Gilst,
 Ruud Van Sloun
Continuous monitoring with an ever-increasing number of sensors has become ubiqu
itous across many application domains. However, acquired time series are typical
ly high-dimensional and difficult to interpret. Expressive deep learning (DL) mo
dels have gained popularity for dimensionality reduction, but the resulting late
nt space often remains difficult to interpret. In this work we propose SOM-CPC,
a model that visualizes data in an organized 2D manifold, while preserving highe
r-dimensional information. We address a largely unexplored and challenging set o
f scenarios comprising high-rate time series, and show on both synthetic and rea
l-life data (physiological data and audio recordings) that SOM-CPC outperforms s
trong baselines like DL-based feature extraction, followed by conventional dimen
sionality reduction techniques, and models that jointly optimize a DL model and
a Self-Organizing Map (SOM). SOM-CPC has great potential to acquire a better und
erstanding of latent patterns in high-rate data streams.
****

One-Shot Federated Conformal Prediction
Pierre Humbert, Batiste Le Bars, Aurélien Bellet, Sylvain Arlot
In this paper, we present a Conformal Prediction method that computes prediction
 sets in a one-shot Federated Learning (FL) setting. More specifically, we intro
duce a novel quantile-of-quantiles estimator and prove that for any distribution
, it is possible to compute prediction sets with desired coverage in only one ro
und of communication. To mitigate privacy issues, we also describe a locally dif
ferentially private version of our estimator. Finally, over a wide range of expe
riments, we show that our method returns prediction sets with coverage and lengt
h very similar to those obtained in a centralized setting. These results demonst
rate that our method is well-suited for one-shot Federated Learning.
****

The Impact of Exploration on Convergence and Performance of Multi-Agent Q-Learni
ng Dynamics
Aamal Hussain, Francesco Belardinelli, Dario Paccagnan
Understanding the impact of exploration on the behaviour of multi-agent learning
 has, so far, benefited from the restriction to potential, or network zero-sum g
ames in which convergence to an equilibrium can be shown. Outside of these class
es, learning dynamics rarely converge and little is known about the effect of ex
ploration in the face of non-convergence. To progress this front, we study the s

mooth Q- Learning dynamics. We show that, in any network game, exploration by ag ents results in the convergence of Q-Learning to a neighbourhood of an equilibri um. This holds independently of whether the dynamics reach the equilibrium or di splay complex behaviours. We show that increasing the exploration rate decreases the size of this neighbourhood and also decreases the ability of all agents to improve their payoffs. Furthermore, in a broad class of games, the payoff perfor mance of Q-Learning dynamics, measured by Social Welfare, decreases when the exp loration rate increases. Our experiments show this to be a general phenomenon, n amely that exploration leads to improved convergence of Q-Learning, at the cost of payoff performance.

****

Combinatorial Neural Bandits

Taehyun Hwang, Kyuwook Chai, Min-Hwan Oh

We consider a contextual combinatorial bandit problem where in each round a lear ning agent selects a subset of arms and receives feedback on the selected arms a ccording to their scores. The score of an arm is an unknown function of the arm’ s feature. Approximating this unknown score function with deep neural networks, we propose algorithms: Combinatorial Neural UCB ($\texttt{CN-UCB}$) and Combinat orial Neural Thompson Sampling ($\texttt{CN-TS}$). We prove that $\texttt{CN-UCB }$ achieves $\tilde{\mathcal{O}}(\tilde{d} \sqrt{T})$ or $\tilde{\mathcal{O}}(\s qrt{\tilde{d} T K})$ regret, where $\tilde{d}$ is the effective dimension of a n eural tangent kernel matrix, K is the size of a subset of arms, and T is the time horizon. For $\texttt{CN-TS}$, we adapt an optimistic sampling technique t o ensure the optimism of the sampled combinatorial action, achieving a worst-cas e (frequentist) regret of $\tilde{\mathcal{O}}(\tilde{d} \sqrt{TK})$. To the bes t of our knowledge, these are the first combinatorial neural bandit algorithms w ith regret performance guarantees. In particular, $\texttt{CN-TS}$ is the first Thompson sampling algorithm with the worst-case regret guarantees for the genera l contextual combinatorial bandit problem. The numerical experiments demonstrate the superior performances of our proposed algorithms.

****

MAGANet: Achieving Combinatorial Generalization by Modeling a Group Action

Geonho Hwang, Jaewoong Choi, Hyunsoo Cho, Myungjoo Kang

Combinatorial generalization refers to the ability to collect and assemble vario us attributes from diverse data to generate novel unexperienced data. This abili ty is considered a necessary passing point for achieving human-level intelligenc e. To achieve this ability, previous unsupervised approaches mainly focused on l earning the disentangled representation, such as the variational autoencoder. Ho wever, recent studies discovered that the disentangled representation is insuffi cient for combinatorial generalization and is not even correlated. In this regar d, we propose a novel framework for data generation that can robustly generalize under these distribution shift situations. Instead of representing each data, o ur model discovers the fundamental transformation between a pair of data by simu lating a group action. To test the combinatorial generalizability, we evaluated our model in two settings: Recombination-to-Element and Recombination-to-Range. The experiments demonstrated that our method has quantitatively and qualitativel y superior generalizability and generates better images than traditional models.

****

Information-Theoretic State Space Model for Multi-View Reinforcement Learning

Hyeongjoo Hwang, Seokin Seo, Youngsoo Jang, Sungyoon Kim, Geon-Hyeong Kim, Seung hoon Hong, Kee-Eung Kim

Multi-View Reinforcement Learning (MVRL) seeks to find an optimal control for an agent given multi-view observations from various sources. Despite recent advanc es in multi-view learning that aim to extract the latent representation from mul ti-view data, it is not straightforward to apply them to control tasks, especial ly when the observations are temporally dependent on one another. The problem ca n be even more challenging if the observations are intermittently missing for a subset of views. In this paper, we introduce Fuse2Control (F2C), an information- theoretic approach to capturing the underlying state space model from the sequen ces of multi-view observations. We conduct an extensive set of experiments in va

rious control tasks showing that our method is highly effective in aggregating t ask-relevant information across many views, that scales linearly with the number of views while retaining robustness to arbitrary missing view scenarios.
****

Under-Counted Tensor Completion with Neural Incorporation of Attributes
Shahana Ibrahim, Xiao Fu, Rebecca Hutchinson, Eugene Seo
Systematic under-counting effects are observed in data collected across many dis ciplines, e.g., epidemiology and ecology. Under-counted tensor completion (UC-TC) is well-motivated for many data analytics tasks, e.g., inferring the case numb ers of infectious diseases at unobserved locations from under-counted case numbe rs in neighboring regions. However, existing methods for similar problems often lack supports in theory, making it hard to understand the underlying principles and conditions beyond empirical successes. In this work, a low-rank Poisson tens or model with an expressive unknown nonlinear side information extractor is prop osed for under-counted multi-aspect data. A joint low-rank tensor completion and neural network learning algorithm is designed to recover the model. Moreover, t he UC-TC formulation is supported by theoretical analysis showing that the fully counted entries of the tensor and each entry’s under-counting probability can b e provably recovered from partial observations—under reasonable conditions. To o ur best knowledge, the result is the first to offer theoretical supports for und er-counted multi-aspect data completion. Simulations and real-data experiments c orroborate the theoretical claims.
****

On the Identifiability and Estimation of Causal Location-Scale Noise Models
Alexander Immer, Christoph Schultheiss, Julia E Vogt, Bernhard Schölkopf, Peter Bühlmann, Alexander Marx
We study the class of location-scale or heteroscedastic noise models (LSNMs), in which the effect Y can be written as a function of the cause X and a noise source N independent of X, which may be scaled by a positive function g ov er the cause, i.e., $Y = f(X) + g(X)N$. Despite the generality of the model clas s, we show the causal direction is identifiable up to some pathological cases. T o empirically validate these theoretical findings, we propose two estimators for LSNMs: an estimator based on (non-linear) feature maps, and one based on neural networks. Both model the conditional distribution of Y given X as a Gaussia n parameterized by its natural parameters. When the feature maps are correctly s pecified, we prove that our estimator is jointly concave, and a consistent estim ator for the cause-effect identification task. Although the the neural network d oes not inherit those guarantees, it can fit functions of arbitrary complexity, and reaches state-of-the-art performance across benchmarks.
****

Stochastic Marginal Likelihood Gradients using Neural Tangent Kernels
Alexander Immer, Tycho F. A. Van Der Ouderaa, Mark Van Der Wilk, Gunnar Ratsch, Bernhard Schölkopf
Selecting hyperparameters in deep learning greatly impacts its effectiveness but requires manual effort and expertise. Recent works show that Bayesian model sel ection with Laplace approximations can allow to optimize such hyperparameters ju st like standard neural network parameters using gradients and on the training d ata. However, estimating a single hyperparameter gradient requires a pass throug h the entire dataset, limiting the scalability of such algorithms. In this work, we overcome this issue by introducing lower bounds to the linearized Laplace ap proximation of the marginal likelihood. In contrast to previous estimators, thes e bounds are amenable to stochastic-gradient-based optimization and allow to tra de off estimation accuracy against computational complexity. We derive them usin g the function-space form of the linearized Laplace, which can be estimated usin g the neural tangent kernel. Experimentally, we show that the estimators can sig nificantly accelerate gradient-based hyperparameter optimization.
****

Differentially Private Hierarchical Clustering with Provable Approximation Guara ntees
Jacob Imola, Alessandro Epasto, Mohammad Mahdian, Vincent Cohen-Addad, Vahab Mir

rokni
Hierarchical Clustering is a popular unsupervised machine learning method with d ecades of history and numerous applications. We initiate the study of differenti ally-private approximation algorithms for hierarchical clustering under the rigo rous framework introduced by Dasgupta (2016). We show strong lower bounds for th e problem: that any ϵ-DP algorithm must exhibit $O(|V|^2/ \epsilon)$-ad ditive error for an input dataset V. Then, we exhibit a polynomial-time approx imation algorithm with $O(|V|^{2.5}/ \epsilon)$-additive error, and an exponenti al-time algorithm that meets the lower bound. To overcome the lower bound, we fo cus on the stochastic block model, a popular model of graphs, and, with a separa tion assumption on the blocks, propose a private $1+o(1)$ approximation algorith m which also recovers the blocks exactly. Finally, we perform an empirical study of our algorithms and validate their performance.
****
Neural Network Accelerated Implicit Filtering: Integrating Neural Network Surrog ates With Provably Convergent Derivative Free Optimization Methods
Brian Irwin, Eldad Haber, Raviv Gal, Avi Ziv
In this paper, we introduce neural network accelerated implicit filtering (NNAIF), a novel family of methods for solving noisy derivative free (i.e. black box, zeroth order) optimization problems. NNAIF intelligently combines the establishe d literature on implicit filtering (IF) optimization methods with a neural netwo rk (NN) surrogate model of the objective function, resulting in accelerated deri vative free methods for unconstrained optimization problems. The NN surrogate mo del consists of a fixed number of parameters, which can be as few as $\approx 1. 3 \times 10^{4}$, that are updated as NNAIF progresses. We show that NNAIF direc tly inherits the convergence properties of IF optimization methods, and thus NNA IF is guaranteed to converge towards a critical point of the objective function under appropriate assumptions. Numerical experiments with 31 noisy problems fr om the CUTEst optimization benchmark set demonstrate the benefits and costs asso ciated with NNAIF. These benefits include NNAIF’s ability to minimize structured functions of several thousand variables much more rapidly than well-known alter natives, such as Covariance Matrix Adaptation Evolution Strategy (CMA-ES) and fi nite difference based variants of gradient descent (GD) and BFGS, as well as its namesake IF.
****
Principled Offline RL in the Presence of Rich Exogenous Information
Riashat Islam, Manan Tomar, Alex Lamb, Yonathan Efroni, Hongyu Zang, Aniket Raji v Didolkar, Dipendra Misra, Xin Li, Harm Van Seijen, Remi Tachet Des Combes, Joh n Langford
Learning to control an agent from offline data collected in a rich pixel-based v isual observation space is vital for real-world applications of reinforcement le arning (RL). A major challenge in this setting is the presence of input informat ion that is hard to model and irrelevant to controlling the agent. This problem has been approached by the theoretical RL community through the lens of exogenou s information, i.e., any control-irrelevant information contained in observation s. For example, a robot navigating in busy streets needs to ignore irrelevant in formation, such as other people walking in the background, textures of objects, or birds in the sky. In this paper, we focus on the setting with visually detail ed exogenous information and introduce new offline RL benchmarks that offer the ability to study this problem. We find that contemporary representation learning techniques can fail on datasets where the noise is a complex and time-dependent process, which is prevalent in practical applications. To address these, we pro pose to use multi-step inverse models to learn Agent-Centric Representations for Offline-RL (ACRO). Despite being simple and reward-free, we show theoretically and empirically that the representation created by this objective greatly outper forms baselines.
****
Unveiling the Latent Space Geometry of Push-Forward Generative Models
Thibaut Issenhuth, Ugo Tanielian, Jeremie Mary, David Picard
Many deep generative models are defined as a push-forward of a Gaussian measure

by a continuous generator, such as Generative Adversarial Networks (GANs) or Var iational Auto-Encoders (VAEs). This work explores the latent space of such deep generative models. A key issue with these models is their tendency to output sam ples outside of the support of the target distribution when learning disconnecte d distributions. We investigate the relationship between the performance of thes e models and the geometry of their latent space. Building on recent developments in geometric measure theory, we prove a sufficient condition for optimality in the case where the dimension of the latent space is larger than the number of mo des. Through experiments on GANs, we demonstrate the validity of our theoretical results and gain new insights into the latent space geometry of these models. A dditionally, we propose a truncation method that enforces a simplicial cluster s tructure in the latent space and improves the performance of GANs.

****

CO-BED: Information-Theoretic Contextual Optimization via Bayesian Experimental Design

Desi R. Ivanova, Joel Jennings, Tom Rainforth, Cheng Zhang, Adam Foster

We formalize the problem of contextual optimization through the lens of Bayesian experimental design and propose CO-BED—a general, model-agnostic framework for designing contextual experiments using information-theoretic principles. After f ormulating a suitable information-based objective, we employ black-box variation al methods to simultaneously estimate it and optimize the designs in a single st ochastic gradient scheme. In addition, to accommodate discrete actions within ou r framework, we propose leveraging continuous relaxation schemes, which can natu rally be integrated into our variational objective. As a result, CO-BED provides a general and automated solution to a wide range of contextual optimization pro blems. We illustrate its effectiveness in a number of experiments, where CO-BED demonstrates competitive performance even when compared to bespoke, model-specif ic alternatives.

****

DoG is SGD’s Best Friend: A Parameter-Free Dynamic Step Size Schedule

Maor Ivgi, Oliver Hinder, Yair Carmon

We propose a tuning-free dynamic SGD step size formula, which we call Distance o ver Gradients (DoG). The DoG step sizes depend on simple empirical quantities (d istance from the initial point and norms of gradients) and have no “learning rat e” parameter. Theoretically, we show that, for stochastic convex optimization, a slight variation of the DoG formula enjoys strong, high-probability parameter-f ree convergence guarantees and iterate movement bounds. Empirically, we consider a broad range of vision and language transfer learning tasks, and show that DoG ’s performance is close to that of SGD with tuned learning rate. We also propose a per-layer variant of DoG that generally outperforms tuned SGD, approaching th e performance of tuned Adam. A PyTorch implementation of our algorithms is avail able at https://github.com/formll/dog.

****

Maximal Initial Learning Rates in Deep ReLU Networks

Gaurav Iyer, Boris Hanin, David Rolnick

Training a neural network requires choosing a suitable learning rate, which invo lves a trade-off between speed and effectiveness of convergence. While there has been considerable theoretical and empirical analysis of how large the learning rate can be, most prior work focuses only on late-stage training. In this work, we introduce the maximal initial learning rate $\eta^{\ast}$ - the largest learn ing rate at which a randomly initialized neural network can successfully begin t raining and achieve (at least) a given threshold accuracy. Using a simple approa ch to estimate $\eta^{\ast}$, we observe that in constant-width fully-connected ReLU networks, $\eta^{\ast}$ behaves differently from the maximum learning rate later in training. Specifically, we find that $\eta^{\ast}$ is well predicted as a power of depth $\times$ width, provided that (i) the width of the network is sufficiently large compared to the depth, and (ii) the input layer is trained at a relatively small learning rate. We further analyze the relationship between $ \eta^{\ast}$ and the sharpness λ_{1} of the network at initialization, i ndicating they are closely though not inversely related. We formally prove bound

s for λ_{1} in terms of depth $\times$ width that align with our empiric
al results.
****
Data-Driven Subgroup Identification for Linear Regression
Zachary Izzo, Ruishan Liu, James Zou
Medical studies frequently require to extract the relationship between each cova
riate and the outcome with statistical confidence measures. To do this, simple p
arametric models are frequently used (e.g. coefficients of linear regression) bu
t always fitted on the whole dataset. However, it is common that the covariates
may not have a uniform effect over the whole population and thus a unified simpl
e model can miss the heterogeneous signal. For example, a linear model may be ab
le to explain a subset of the data but fail on the rest due to the nonlinearity
and heterogeneity in the data. In this paper, we propose DDGroup (data-driven gr
oup discovery), a data-driven method to effectively identify subgroups in the da
ta with a uniform linear relationship between the features and the label. DDGrou
p outputs an interpretable region in which the linear model is expected to hold.
It is simple to implement and computationally tractable for use. We show theore
tically that, given a large enough sample, DDGroup recovers a region where a sin
gle linear model with low variance is well-specified (if one exists), and experi
ments on real-world medical datasets confirm that it can discover regions where
a local linear model has improved performance. Our experiments also show that DD
Group can uncover subgroups with qualitatively different relationships which are
missed by simply applying parametric approaches to the whole dataset.
****
Efficient Training of Language Models using Few-Shot Learning
Sashank J. Reddi, Sobhan Miryoosefi, Stefani Karp, Shankar Krishnan, Satyen Kale
, Seungyeon Kim, Sanjiv Kumar
Large deep learning models have achieved state-of-the-art performance across var
ious natural language processing (NLP) tasks and demonstrated remarkable few-sho
t learning performance. However, training them is often challenging and resource
-intensive. In this paper, we study an efficient approach to train language mode
ls using few-shot learners. We show that, by leveraging the fast learning nature
of few-shot learners, one can train language models efficiently in a stagewise
manner. Our main insight is that stacking a good few-shot learner on a good smal
l language model provides a good initializer for a larger language model. Using
this insight and building upon progressive stacking approaches, we develop novel
approaches for training such networks in a stagewise manner. Furthermore, we al
so provide a theoretical framework and accompanying empirical studies to support
our insights, thereby creating a theoretical foundation for progressive stackin
g. Finally, we provide empirical results to demonstrate the effectiveness of our
approach in reducing the training time of few-shot learners.
****
Scalable Adaptive Computation for Iterative Generation
Allan Jabri, David J. Fleet, Ting Chen
Natural data is redundant yet predominant architectures tile computation uniform
ly across their input and output space. We propose the Recurrent Interface Netwo
rk (RIN), an attention-based architecture that decouples its core computation fr
om the dimensionality of the data, enabling adaptive computation for more scalab
le generation of high-dimensional data. RINs focus the bulk of computation (i.e.
global self-attention) on a set of latent tokens, using cross-attention to read
and write (i.e. route) information between latent and data tokens. Stacking RIN
blocks allows bottom-up (data to latent) and top-down (latent to data) feedback
, leading to deeper and more expressive routing. While this routing introduces c
hallenges, this is less problematic in recurrent computation settings where the
task (and routing problem) changes gradually, such as iterative generation with
diffusion models. We show how to leverage recurrence by conditioning the latent
tokens at each forward pass of the reverse diffusion process with those from pri
or computation, i.e. latent self-conditioning. RINs yield state-of-the-art pixel
diffusion models for image and video generation, scaling to1024×1024 images wit
hout cascades or guidance, while being domain-agnostic and up to 10× more effici

ent than 2D and 3D U-Nets.
****

Unconstrained Online Learning with Unbounded Losses
Andrew Jacobsen, Ashok Cutkosky
Algorithms for online learning typically require one or more boundedness assumpt ions: that the domain is bounded, that the losses are Lipschitz, or both. In thi s paper, we develop a new setting for online learning with unbounded domains and non-Lipschitz losses. For this setting we provide an algorithm which guarantees $R_{T}(u)\le \tilde O(G\|u\|\sqrt{T}+L\|u\|^{2}\sqrt{T})$ regret on any problem where the subgradients satisfy $\|g_{t}\|\le G+L\|w_{t}\|$, and show that this bound is unimprovable without further assumptions. We leverage this algorithm to develop new saddle-point optimization algorithms that converge in duality gap i n unbounded domains, even in the absence of meaningful curvature. Finally, we pr ovide the first algorithm achieving non-trivial dynamic regret in an unbounded d omain for non-Lipschitz losses, as well as a matching lower bound. The regret of our dynamic regret algorithm automatically improves to a novel L^{*} bound wh en the losses are smooth.
****

Multi-Objective GFlowNets
Moksh Jain, Sharath Chandra Raparthy, Alex Hernández-Garc■■a, Jarrid Rector-Broo ks, Yoshua Bengio, Santiago Miret, Emmanuel Bengio
We study the problem of generating diverse candidates in the context of Multi-Ob jective Optimization. In many applications of machine learning such as drug disc overy and material design, the goal is to generate candidates which simultaneous ly optimize a set of potentially conflicting objectives. Moreover, these objecti ves are often imperfect evaluations of some underlying property of interest, mak ing it important to generate diverse candidates to have multiple options for exp ensive downstream evaluations. We propose Multi-Objective GFlowNets (MOGFNs), a novel method for generating diverse Pareto optimal solutions, based on GFlowNets . We introduce two variants of MOGFNs: MOGFN-PC, which models a family of indepe ndent sub-problems defined by a scalarization function, with reward-conditional GFlowNets, and MOGFN-AL, which solves a sequence of sub-problems defined by an a cquisition function in an active learning loop. Our experiments on wide variety of synthetic and benchmark tasks demonstrate advantages of the proposed methods in terms of the Pareto performance and importantly, improved candidate diversity , which is the main contribution of this work.
****

The Price of Differential Privacy under Continual Observation
Palak Jain, Sofya Raskhodnikova, Satchit Sivakumar, Adam Smith
We study the accuracy of differentially private mechanisms in the continual rele ase model. A continual release mechanism receives a sensitive dataset as a strea m of T inputs and produces, after receiving each input, an output that is accu rate for all the inputs received so far. We provide the first strong lower bound s on the error of continual release mechanisms. In particular, for two fundament al problems that are closely related to empirical risk minimization and widely s tudied and used in the standard (batch) model, we prove that the worst case erro r of every continual release algorithm is $\tilde \Omega(T^{1/3})$ times larger than that of the best batch algorithm. Previous work shows only a $\Omega(\log T)$ gap between the worst case error achievable in these two models. We also form ulate a model that allows for adaptively selected inputs, thus capturing depende ncies that arise in many applications of continual release. Even though, in gene ral, both privacy and accuracy are harder to attain in this model, we show that our lower bounds are matched by the error of simple algorithms that work even fo r adaptively selected inputs.
****

Graph Ladling: Shockingly Simple Parallel GNN Training without Intermediate Comm unication
Ajay Kumar Jaiswal, Shiwei Liu, Tianlong Chen, Ying Ding, Zhangyang Wang
Graphs are omnipresent and GNNs are a powerful family of neural networks for lea rning over graphs. Despite their popularity, scaling GNNs either by deepening or

widening suffers from prevalent issues of $\textit{unhealthy gradients, over-smoothening, information squashing}$, which often lead to sub-standard performance. In this work, we are interested in exploring a principled way to scale GNNs capacity without deepening or widening, which can improve its performance across multiple small and large graphs. Motivated by the recent intriguing phenomenon of model soups, which suggest that fine-tuned weights of multiple large-language pre-trained models can be merged to a better minima, we argue to exploit the fundamentals of model soups to mitigate the aforementioned issues of memory bottleneck and trainability during GNNs scaling. More specifically, we propose not to deepen or widen current GNNs, but instead present $\textbf{first data-centric perspective}$ of model soups to build powerful GNNs by dividing giant graph data to build independently and parallelly trained multiple comparatively weaker GNNs without any intermediate communication, and $\textit{combining their strength}$ using a greedy interpolation soup procedure to achieve state-of-the-art performance. Moreover, we provide a wide variety of model soup preparation techniques by leveraging state-of-the-art graph sampling and graph partitioning approaches that can handle large graph data structures. Our extensive experiments across many real-world small and large graphs, illustrate the effectiveness of our approach and point towards a promising orthogonal direction for GNN scaling. Codes are available at: https://github.com/VITA-Group/graph_ladling
****

Instant Soup: Cheap Pruning Ensembles in A Single Pass Can Draw Lottery Tickets from Large Models

Ajay Kumar Jaiswal, Shiwei Liu, Tianlong Chen, Ying Ding, Zhangyang Wang

Large pre-trained transformers have been receiving explosive attention in the past few years, due to their acculturation for numerous downstream applications via fine-tuning, but their exponentially increasing parameter counts are becoming a primary hurdle to even just fine-tune them without industry-standard hardware. Recently, Lottery Ticket Hypothesis (LTH) and its variants, have been exploited to prune these large pre-trained models generating subnetworks which can achieve similar performance as their dense counterparts, but LTH pragmatism is enormously inhibited by repetitive full training and pruning routine of iterative magnitude pruning (IMP) which worsens with increasing model size. Motivated by the recent observations of model soups, which suggest that fine-tuned weights of multiple models can be merged to a better minima, we propose Instant Soup Pruning (ISP) to generate lottery ticket quality subnetworks, using a fraction of the original IMP cost by replacing the expensive intermediate pruning stages of IMP with computationally efficient weak mask generation and aggregation routine. More specifically, during the mask generation stage, ISP takes a small handful of iterations using varying training protocols and data subsets to generate many weak and noisy subnetworks, and superpose them to average out the noise creating a high-quality denoised subnetwork. Our extensive experiments and ablation on two popular large-scale pre-trained models: $\texttt{CLIP} (unexplored in pruning till date)$ and $\texttt{BERT}$ across multiple benchmark vision $\texttt{\{MNIST, SVHN, Cars, GTSRB, CIFAR-10, CIFAR-100\}}$ and language datasets $\texttt{\{MNLI, QNLI, QQP, SST, ...\}}$ validate the effectiveness of ISP compared to several state-of-the-art pruning methods. Additionally, we show that ISP can be easily modified with minimal overhead to produce benefits comparable to model soups, without the prerequisite to generate multiple candidates fine-tuned models. Codes are available at: https://github.com/VITA-Group/instant_soup.
****

Exploring the Benefits of Training Expert Language Models over Instruction Tuning

Joel Jang, Seungone Kim, Seonghyeon Ye, Doyoung Kim, Lajanugen Logeswaran, Moontae Lee, Kyungjae Lee, Minjoon Seo

Recently, Language Models (LMs) instruction-tuned on multiple tasks, also known as multitask-prompted fine-tuning (MT), have shown capabilities to generalize to unseen tasks. Previous work has shown that scaling the number of finetuning datasets and instructions is the key component in making stronger MT LMs. In this work, we report surprising findings that show an expert LM trained on just a sing

le task can outperform an MT LM trained with 300+ different tasks on 11 differen t unseen datasets and on 13 datasets of the BIG-bench benchmark by an average of 3.20% and 1.29%, respectively. This finding casts doubt on the previously held belief that simply scaling the number of tasks makes stronger MT LMs. Leveraging this finding, we further show that this distributed approach of training multip le expert LMs instead of a single MT LM for zero-shot inference possesses many b enefits including (1) avoiding negative task transfer that often occurs during i nstruction tuning, (2) being able to continually learn new tasks without having to re-train on previous tasks to avoid catastrophic forgetting, and (3) showing compositional capabilities when merging individual experts together.
****

Learning to Boost Training by Periodic Nowcasting Near Future Weights
Jinhyeok Jang, Woo-Han Yun, Won Hwa Kim, Youngwoo Yoon, Jaehong Kim, Jaeyeon Lee , Byungok Han
Recent complicated problems require large-scale datasets and complex model archi tectures, however, it is difficult to train such large networks due to high comp utational issues. Significant efforts have been made to make the training more e fficient such as momentum, learning rate scheduling, weight regularization, and meta-learning. Based on our observations on 1) high correlation between past eig hts and future weights, 2) conditions for beneficial weight prediction, and 3) f easibility of weight prediction, we propose a more general framework by intermit tently skipping a handful of epochs by periodically forecasting near future weig hts, i.e., a Weight Nowcaster Network (WNN). As an add-on module, WNN predicts t he future weights to make the learning process faster regardless of tasks and ar chitectures. Experimental results show that WNN can significantly save actual ti me cost for training with an additional marginal time to train WNN. We validate the generalization capability of WNN under various tasks, and demonstrate that i t works well even for unseen tasks. The code and pre-trained model are available at https://github.com/jjh6297/WNN.
****

Unscented Autoencoder
Faris Janjos, Lars Rosenbaum, Maxim Dolgov, J. Marius Zoellner
The Variational Autoencoder (VAE) is a seminal approach in deep generative model ing with latent variables. Interpreting its reconstruction process as a nonlinea r transformation of samples from the latent posterior distribution, we apply the Unscented Transform (UT) – a well-known distribution approximation used in the Unscented Kalman Filter (UKF) from the field of filtering. A finite set of stati stics called sigma points, sampled deterministically, provides a more informativ e and lower-variance posterior representation than the ubiquitous noise-scaling of the reparameterization trick, while ensuring higher-quality reconstruction. W e further boost the performance by replacing the Kullback-Leibler (KL) divergenc e with the Wasserstein distribution metric that allows for a sharper posterior. Inspired by the two components, we derive a novel, deterministic-sampling flavor of the VAE, the Unscented Autoencoder (UAE), trained purely with regularization -like terms on the per-sample posterior. We empirically show competitive perform ance in Fréchet Inception Distance scores over closely-related models, in additi on to a lower training variance than the VAE.
****

Curiosity in Hindsight: Intrinsic Exploration in Stochastic Environments
Daniel Jarrett, Corentin Tallec, Florent Altché, Thomas Mesnard, Remi Munos, Mic hal Valko
Consider the problem of exploration in sparse-reward or reward-free environments , such as in Montezuma’s Revenge. In the curiosity-driven paradigm, the agent is rewarded for how much each realized outcome differs from their predicted outcom e. But using predictive error as intrinsic motivation is fragile in stochastic e nvironments, as the agent may become trapped by high-entropy areas of the state- action space, such as a "noisy TV". In this work, we study a natural solution de rived from structural causal models of the world: Our key idea is to learn repre sentations of the future that capture precisely the unpredictable aspects of eac h outcome—which we use as additional input for predictions, such that intrinsic

rewards only reflect the predictable aspects of world dynamics. First, we propos e incorporating such hindsight representations into models to disentangle "noise " from "novelty", yielding Curiosity in Hindsight: a simple and scalable general ization of curiosity that is robust to stochasticity. Second, we instantiate thi s framework for the recently introduced BYOL-Explore algorithm as our prime exam ple, resulting in the noise-robust BYOL-Hindsight. Third, we illustrate its beha vior under a variety of different stochasticities in a grid world, and find impr ovements over BYOL-Explore in hard-exploration Atari games with sticky actions. Notably, we show state-of-the-art results in exploring Montezuma’s Revenge with sticky actions, while preserving performance in the non-sticky setting.
****

BiRT: Bio-inspired Replay in Vision Transformers for Continual Learning
Kishaan Jeeveswaran, Prashant Shivaram Bhat, Bahram Zonooz, Elahe Arani
The ability of deep neural networks to continually learn and adapt to a sequence of tasks has remained challenging due to catastrophic forgetting of previously learned tasks. Humans, on the other hand, have a remarkable ability to acquire, assimilate, and transfer knowledge across tasks throughout their lifetime withou t catastrophic forgetting. The versatility of the brain can be attributed to the rehearsal of abstract experiences through a complementary learning system. Howe ver, representation rehearsal in vision transformers lacks diversity, resulting in overfitting and consequently, performance drops significantly compared to raw image rehearsal. Therefore, we propose BiRT, a novel representation rehearsal-b ased continual learning approach using vision transformers. Specifically, we int roduce controllable noises at various stages of the vision transformer and enfor ce consistency in predictions with respect to an exponential moving average of t he working model. Our method provides consistent performance gain over raw image and vanilla representation rehearsal on several challenging CL benchmarks while being memory efficient and robust to natural and adversarial corruptions.
****

Recovering Top-Two Answers and Confusion Probability in Multi-Choice Crowdsourci ng
Hyeonsu Jeong, Hye Won Chung
Crowdsourcing has emerged as an effective platform for labeling large amounts of data in a cost- and time-efficient manner. Most previous work has focused on de signing an efficient algorithm to recover only the ground-truth labels of the da ta. In this paper, we consider multi-choice crowdsourcing tasks with the goal of recovering not only the ground truth, but also the most confusing answer and th e confusion probability. The most confusing answer provides useful information a bout the task by revealing the most plausible answer other than the ground truth and how plausible it is. To theoretically analyze such scenarios, we propose a model in which there are the top two plausible answers for each task, distinguis hed from the rest of the choices. Task difficulty is quantified by the probabili ty of confusion between the top two, and worker reliability is quantified by the probability of giving an answer among the top two. Under this model, we propose a two-stage inference algorithm to infer both the top two answers and the confu sion probability. We show that our algorithm achieves the minimax optimal conver gence rate. We conduct both synthetic and real data experiments and demonstrate that our algorithm outperforms other recent algorithms. We also show the applica bility of our algorithms in inferring the difficulty of tasks and in training ne ural networks with top-two soft labels.
****

Leveraging Label Non-Uniformity for Node Classification in Graph Neural Networks
Feng Ji, See Hian Lee, Hanyang Meng, Kai Zhao, Jielong Yang, Wee Peng Tay
In node classification using graph neural networks (GNNs), a typical model gener ates logits for different class labels at each node. A softmax layer often outpu ts a label prediction based on the largest logit. We demonstrate that it is poss ible to infer hidden graph structural information from the dataset using these l ogits. We introduce the key notion of label non-uniformity, which is derived fro m the Wasserstein distance between the softmax distribution of the logits and th e uniform distribution. We demonstrate that nodes with small label non-uniformit

y are harder to classify correctly. We theoretically analyze how the label non-uniformity varies across the graph, which provides insights into boosting the model performance: increasing training samples with high non-uniformity or dropping edges to reduce the maximal cut size of the node set of small non-uniformity. These mechanisms can be easily added to a base GNN model. Experimental results demonstrate that our approach improves the performance of many benchmark base models.
****
Bidirectional Adaptation for Robust Semi-Supervised Learning with Inconsistent Data Distributions
Lin-Han Jia, Lan-Zhe Guo, Zhi Zhou, Jie-Jing Shao, Yuke Xiang, Yu-Feng Li
Semi-supervised learning (SSL) suffers from severe performance degradation when labeled and unlabeled data come from inconsistent data distributions. However, there is still a lack of sufficient theoretical guidance on how to alleviate this problem. In this paper, we propose a general theoretical framework that demonstrates how distribution discrepancies caused by pseudo-label predictions and target predictions can lead to severe generalization errors. Through theoretical analysis, we identify three main reasons why previous SSL algorithms cannot perform well with inconsistent distributions: coupling between the pseudo-label predictor and the target predictor, biased pseudo labels, and restricted sample weights. To address these challenges, we introduce a practical framework called Bidirectional Adaptation that can adapt to the distribution of unlabeled data for debiased pseudo-label prediction and to the target distribution for debiased target prediction, thereby mitigating these shortcomings. Extensive experimental results demonstrate the effectiveness of our proposed framework.
****
Short-lived High-volume Bandits
Su Jia, Nishant Oli, Ian Anderson, Paul Duff, Andrew A Li, R. Ravi
Modern platforms leverage randomized experiments to make informed decisions from a given set of alternatives. As a particularly challenging scenario, these alternatives can potentially have (i) high volume, with thousands of new items being released each hour, and (ii) short lifetime, either due to the contents' transient nature, or some underlying non-stationarity that impels the learner to treat the same item as non-identical copies across time. We consider a multiplay bandits model. In each round a set of $k=n^\rho$ actions that will be available for w rounds arrives, each of whose mean reward is drawn from a fixed known distribution. The learner selects a multiset of n actions at a time. We propose an ℓ-Layered Sieve Policy that recursively refines the action space for $\ell\leq w$ times. We show that for any given $\rho>0$, with suitable ℓ, the policy achieves $\tilde O (n^{-\min \{\rho, \frac 12 (1+\frac 1w)^{-1}\}})$ regret. We also complement this result with an $\Omega (n^{-\min \{\rho, \frac 12\}})$ lower bound. We further validate the effectiveness of our Sieve Policy via numerical simulations and a field experiment in a large content card serving platform.
****
Smooth Non-stationary Bandits
Su Jia, Qian Xie, Nathan Kallus, Peter I. Frazier
In many applications of online decision making, the environment is non-stationary and it is therefore crucial to use bandit algorithms that handle changes. Most existing approaches are designed to protect against non-smooth changes, constrained only by total variation or Lipschitzness over time, where they guarantee $T^{2/3}$ regret. However, in practice environments are often changing smoothly, so such algorithms may incur higher-than-necessary regret in these settings and do not leverage information on the rate of change. In this paper, we study a non-stationary two-arm bandit problem where we assume an arm's mean reward is a β-Hölder function over (normalized) time, meaning it is $(\beta-1)$-times Lipschitz-continuously differentiable. We show the first separation between the smooth and non-smooth regimes by presenting a policy with $T^{3/5}$ regret for $\beta=2$. We complement this result by a $T^{\frac{\beta+1}{2\beta+1}}$ lower bound for any integer $\beta\ge 1$, which matches our upper bound for $\beta=2$.
****

A Unified Optimization Framework of ANN-SNN Conversion: Towards Optimal Mapping from Activation Values to Firing Rates
Haiyan Jiang, Srinivas Anumasa, Giulia De Masi, Huan Xiong, Bin Gu
Spiking Neural Networks (SNNs) have gained significant attention for their energy-efficient and fast-inference capabilities, but training SNNs from scratch can be challenging due to the discrete nature of spikes. One alternative method is to convert an Artificial Neural Network (ANN) into an SNN, known as ANN-SNN conversion. Currently, existing ANN-SNN conversion methods often involve redesigning the ANN with a new activation function, rather than utilizing the traditional ReLU, and converting it to an SNN. However, these methods do not take into account the potential performance loss between the regular ANN with ReLU and the tailored ANN. In this work, we propose a unified optimization framework for ANN-SNN conversion that considers both performance loss and conversion error. To achieve this, we introduce the SlipReLU activation function, which is a weighted sum of the threshold-ReLU and the step function. Theoretical analysis demonstrates that conversion error can be zero on a range of shift values $\delta \in [-0.5,0.5]$ rather than a fixed shift term 0.5. We evaluate our SlipReLU method on CIFAR datasets, which shows that SlipReLU outperforms current ANN-SNN conversion methods and supervised training methods in terms of accuracy and latency. To the best of our knowledge, this is the first ANN-SNN conversion method that enables SNN inference using only 1 time step. Code is available at https://github.com/HaiyanJiang/SNN_Conversion_unified.
****
VIMA: Robot Manipulation with Multimodal Prompts
Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, Linxi Fan
Prompt-based learning has emerged as a successful paradigm in natural language processing, where a single general-purpose language model can be instructed to perform any task specified by input prompts. Yet task specification in robotics comes in various forms, such as imitating one-shot demonstrations, following language instructions, and reaching visual goals. They are often considered different tasks and tackled by specialized models. We show that a wide spectrum of robot manipulation tasks can be expressed with multimodal prompts, interleaving textual and visual tokens. Accordingly, we develop a new simulation benchmark that consists of thousands of procedurally-generated tabletop tasks with multimodal prompts, 600K+ expert trajectories for imitation learning, and a four-level evaluation protocol for systematic generalization. We design a transformer-based robot agent, VIMA, that processes these prompts and outputs motor actions autoregressively. VIMA features a recipe that achieves strong model scalability and data efficiency. It outperforms alternative designs in the hardest zero-shot generalization setting by up to $2.9\times$ task success rate given the same training data. With $10\times$ less training data, VIMA still performs $2.7\times$ better than the best competing variant. Code and video demos are available at https://vimalabs.github.io
****
Estimating Causal Effects using a Multi-task Deep Ensemble
Ziyang Jiang, Zhuoran Hou, Yiling Liu, Yiman Ren, Keyu Li, David Carlson
A number of methods have been proposed for causal effect estimation, yet few have demonstrated efficacy in handling data with complex structures, such as images. To fill this gap, we propose Causal Multi-task Deep Ensemble (CMDE), a novel framework that learns both shared and group-specific information from the study population. We provide proofs demonstrating equivalency of CDME to a multi-task Gaussian process (GP) with a coregionalization kernel a priori. Compared to multi-task GP, CMDE efficiently handles high-dimensional and multi-modal covariates and provides pointwise uncertainty estimates of causal effects. We evaluate our method across various types of datasets and tasks and find that CMDE outperforms state-of-the-art methods on a majority of these tasks.
****
Online Restless Bandits with Unobserved States
Bowen Jiang, Bo Jiang, Jian Li, Tao Lin, Xinbing Wang, Chenghu Zhou

We study the online restless bandit problem, where each arm evolves according to a Markov chain independently, and the reward of pulling an arm depends on both the current state of the corresponding Markov chain and the pulled arm. The agent (decision maker) does not know the transition functions and reward functions, and cannot observe the states of arms even after pulling. The goal is to sequentially choose which arms to pull so as to maximize the expected cumulative rewards collected. In this paper, we propose TSEETC, a learning algorithm based on Thompson Sampling with Episodic Explore-Then-Commit. The algorithm proceeds in episodes of increasing length and each episode is divided into exploration and exploitation phases. During the exploration phase, samples of action-reward pairs are collected in a round-robin fashion and utilized to update the posterior distribution as a mixture of Dirichlet distributions. At the beginning of the exploitation phase, TSEETC generates a sample from the posterior distribution as true parameters. It then follows the optimal policy for the sampled model for the rest of the episode. We establish the Bayesian regret bound $\tilde {\mathcal{O}}(\sqrt{T})$ for TSEETC, where T is the time horizon. We show through simulations that TSEETC outperforms existing algorithms in regret.
****

Detecting Out-of-distribution Data through In-distribution Class Prior
Xue Jiang, Feng Liu, Zhen Fang, Hong Chen, Tongliang Liu, Feng Zheng, Bo Han
Given a pre-trained in-distribution (ID) model, the inference-time out-of-distribution (OOD) detection aims to recognize OOD data during the inference stage. However, some representative methods share an unproven assumption that the probability that OOD data belong to every ID class should be the same, i.e., these OOD-to-ID probabilities actually form a uniform distribution. In this paper, we show that this assumption makes the above methods incapable when the ID model is trained with class-imbalanced data.Fortunately, by analyzing the causal relations between ID/OOD classes and features, we identify several common scenarios where the OOD-to-ID probabilities should be the ID-class-prior distribution and propose two strategies to modify existing inference-time detection methods: 1) replace the uniform distribution with the ID-class-prior distribution if they explicitly use the uniform distribution; 2) otherwise, reweight their scores according to the similarity between the ID-class-prior distribution and the softmax outputs of the pre-trained model. Extensive experiments show that both strategies can improve the OOD detection performance when the ID model is pre-trained with imbalanced data, reflecting the importance of ID-class prior in OOD detection.
****

Towards Stable and Efficient Adversarial Training against l_1 Bounded Adversarial Attacks
Yulun Jiang, Chen Liu, Zhichao Huang, Mathieu Salzmann, Sabine Susstrunk
We address the problem of stably and efficiently training a deep neural network robust to adversarial perturbations bounded by an l_1 norm. We demonstrate that achieving robustness against l_1-bounded perturbations is more challenging than in the l_2 or l_∞ cases, because adversarial training against l_1-bounded perturbations is more likely to suffer from catastrophic overfitting and yield training instabilities. Our analysis links these issues to the coordinate descent strategy used in existing methods. We address this by introducing Fast-EG-l_1, an efficient adversarial training algorithm based on Euclidean geometry and free of coordinate descent. Fast-EG-l_1 comes with no additional memory costs and no extra hyper-parameters to tune. Our experimental results on various datasets demonstrate that Fast-EG-l_1 yields the best and most stable robustness against l_1-bounded adversarial attacks among the methods of comparable computational complexity. Code and the checkpoints are available at https://github.com/IVRL/FastAdvL.
****

Learning Unnormalized Statistical Models via Compositional Optimization
Wei Jiang, Jiayu Qin, Lingyu Wu, Changyou Chen, Tianbao Yang, Lijun Zhang
Learning unnormalized statistical models (e.g., energy-based models) is computationally challenging due to the complexity of handling the partition function. To eschew this complexity, noise-contrastive estimation (NCE) has been proposed by

formulating the objective as the logistic loss of the real data and the artific
ial noise. However, as found in previous works, NCE may perform poorly in many t
asks due to its flat loss landscape and slow convergence. In this paper, we stud
y a direct approach for optimizing the negative log-likelihood of unnormalized m
odels from the perspective of compositional optimization. To tackle the partitio
n function, a noise distribution is introduced such that the log partition funct
ion can be written as a compositional function whose inner function can be estim
ated with stochastic samples. Hence, the objective can be optimized by stochasti
c compositional optimization algorithms. Despite being a simple method, we demon
strate that it is more favorable than NCE by (1) establishing a fast convergence
rate and quantifying its dependence on the noise distribution through the varia
nce of stochastic estimators; (2) developing better results for one-dimensional
Gaussian mean estimation by showing our objective has a much favorable loss land
scape and hence our method enjoys faster convergence; (3) demonstrating better p
erformance on multiple applications, including density estimation, out-of-distri
bution detection, and real image generation.
****

Approximate Causal Effect Identification under Weak Confounding
Ziwei Jiang, Lai Wei, Murat Kocaoglu
Causal effect estimation has been studied by many researchers when only observat
ional data is available. Sound and complete algorithms have been developed for p
ointwise estimation of identifiable causal queries. For non-identifiable causal
queries, researchers developed polynomial programs to estimate tight bounds on c
ausal effect. However, these are computationally difficult to optimize for varia
bles with large support sizes. In this paper, we analyze the effect of "weak con
founding’" on causal estimands. More specifically, under the assumption that the
unobserved confounders that render a query non-identifiable have small entropy,
we propose an efficient linear program to derive the upper and lower bounds of
the causal effect. We show that our bounds are consistent in the sense that as t
he entropy of unobserved confounders goes to zero, the gap between the upper and
lower bound vanishes. Finally, we conduct synthetic and real data simulations t
o compare our bounds with the bounds obtained by the existing work that cannot i
ncorporate such entropy constraints and show that our bounds are tighter for the
setting with weak confounders.
****

MEWL: Few-shot multimodal word learning with referential uncertainty
Guangyuan Jiang, Manjie Xu, Shiji Xin, Wei Liang, Yujia Peng, Chi Zhang, Yixin Z
hu
Without explicit feedback, humans can rapidly learn the meaning of words. Childr
en can acquire a new word after just a few passive exposures, a process known as
fast mapping. This word learning capability is believed to be the most fundamen
tal building block of multimodal understanding and reasoning. Despite recent adv
ancements in multimodal learning, a systematic and rigorous evaluation is still
missing for human-like word learning in machines. To fill in this gap, we introd
uce the MachinE Word Learning (MEWL) benchmark to assess how machines learn word
meaning in grounded visual scenes. MEWL covers human’s core cognitive toolkits
in word learning: cross-situational reasoning, bootstrapping, and pragmatic lear
ning. Specifically, MEWL is a few-shot benchmark suite consisting of nine tasks
for probing various word learning capabilities. These tasks are carefully design
ed to be aligned with the children’s core abilities in word learning and echo th
e theories in the developmental literature. By evaluating multimodal and unimoda
l agents’ performance with a comparative analysis of human performance, we notic
e a sharp divergence in human and machine word learning. We further discuss thes
e differences between humans and machines and call for human-like few-shot word
learning in machines.
****

NeuralSlice: Neural 3D Triangle Mesh Reconstruction via Slicing 4D Tetrahedral M
eshes
Chenbo Jiang, Jie Yang, Shwai He, Yu-Kun Lai, Lin Gao
Learning-based high-fidelity reconstruction of 3D shapes with varying topology i

s a fundamental problem in computer vision and computer graphics. Recent advance
s in learning 3D shapes using explicit and implicit representations have achieve
d impressive results in 3D modeling. However, the template-based explicit repres
entation is limited by fixed topology, and the implicit representation, although
 flexible with arbitrary topology, requires a large number of sampled points to
regress the surface, which is computationally expensive. In this work, we propos
e a novel 3D shape representation named NeuralSlice, which represents a 3D shape
 as the intersection of a 4D tetrahedral mesh and a 4D hyperplane. A novel netwo
rk is designed to incorporate the proposed representation flexibly, which learns
 a deformable 4D template and a parameter for slicing 4D hyperplane to reconstru
ct the 3D object. To learn the local deformation of the 4D template, we further
propose a spatial-aware network to locate the 4D points within the 3D feature vo
lume of input shape via positional encoding, which leverages the local geometric
al feature to guide the 4D deformation. By addressing the 3D problem in a higher
 4D space, our method supports flexible topology changes while being highly effi
cient. Our method is guaranteed to produce manifold meshes. NeuralSlice outperfo
rms the state-of-the-art explicit-based approaches in terms of reconstruction qu
ality. Compared with implicit approaches, by avoiding point sampling, our method
 is 10 times faster than the implicit approaches, and better preserves thin stru
ctures. NeuralSlice has the capability of representing various shapes and topolo
gies using a single 4D tetrahedral mesh. The corresponding code can be found on
GitHub at https://github.com/IGLICT/NEURALSLICE
****

Effective Structured Prompting by Meta-Learning and Representative Verbalizer
Weisen Jiang, Yu Zhang, James Kwok
Prompt tuning for pre-trained masked language models (MLM) has shown promising p
erformance in natural language processing tasks with few labeled examples. It tu
nes a prompt for the downstream task, and a verbalizer is used to bridge the pre
dicted token and label prediction. Due to the limited training data, prompt init
ialization is crucial for prompt tuning. Recently, MetaPrompting (Hou et al., 20
22) uses meta-learning to learn a shared initialization for all task-specific pr
ompts. However, a single initialization is insufficient to obtain good prompts f
or all tasks and samples when the tasks are complex. Moreover, MetaPrompting req
uires tuning the whole MLM, causing a heavy burden on computation and memory as
the MLM is usually large. To address these issues, we use a prompt pool to extra
ct more task knowledge and construct instance-dependent prompts via attention. W
e further propose a novel soft verbalizer (RepVerb) which constructs label embed
ding from feature embeddings directly. Combining meta-learning the prompt pool a
nd RepVerb, we propose MetaPrompter for effective structured prompting. MetaProm
pter is parameter-efficient as only the pool is required to be tuned. Experiment
al results demonstrate that MetaPrompter performs better than the recent state-o
f-the-arts and RepVerb outperforms existing soft verbalizers.
****

Understanding Incremental Learning of Gradient Descent: A Fine-grained Analysis
of Matrix Sensing
Jikai Jin, Zhiyuan Li, Kaifeng Lyu, Simon Shaolei Du, Jason D. Lee
It is believed that Gradient Descent (GD) induces an implicit bias towards good
generalization in training machine learning models. This paper provides a fine-g
rained analysis of the dynamics of GD for the matrix sensing problem, whose goal
 is to recover a low-rank ground-truth matrix from near-isotropic linear measure
ments. It is shown that GD with small initialization behaves similarly to the gr
eedy low-rank learning heuristics and follows an incremental learning procedure:
 GD sequentially learns solutions with increasing ranks until it recovers the gr
ound truth matrix. Compared to existing works which only analyze the first learn
ing phase for rank-1 solutions, our result provides characterizations for the wh
ole learning process. Moreover, besides the over-parameterized regime that many
prior works focused on, our analysis of the incremental learning procedure also
applies to the under-parameterized regime. Finally, we conduct numerical experim
ents to confirm our theoretical findings.
****

Thompson Sampling with Less Exploration is Fast and Optimal
Tianyuan Jin, Xianglin Yang, Xiaokui Xiao, Pan Xu
We propose ϵ-Exploring Thompson Sampling (ϵ-TS), a modified version of the Thompson Sampling (TS) algorithm for multi-armed bandits. In ϵ-TS, arms are selected greedily based on empirical mean rewards with probability $1-\epsilon$, and based on posterior samples obtained from TS with probability ϵ. Here, $\epsilon\in(0,1)$ is a user-defined constant. By reducing exploration, ϵ-TS improves computational efficiency compared to TS while achieving better regret bounds. We establish that ϵ-TS is both minimax optimal and asymptotically optimal for various popular reward distributions, including Gaussian, Bernoulli, Poisson, and Gamma. A key technical advancement in our analysis is the relaxation of the requirement for a stringent anti-concentration bound of the posterior distribution, which was necessary in recent analyses that achieved similar bounds. As a result, ϵ-TS maintains the posterior update structure of TS while minimizing alterations, such as clipping the sampling distribution or solving the inverse of the Kullback-Leibler (KL) divergence between reward distributions, as done in previous work. Furthermore, our algorithm is as easy to implement as TS, but operates significantly faster due to reduced exploration. Empirical evaluations confirm the efficiency and optimality of ϵ-TS.
****
R-U-SURE? Uncertainty-Aware Code Suggestions By Maximizing Utility Across Random User Intents
Daniel D. Johnson, Daniel Tarlow, Christian Walder
Large language models show impressive results at predicting structured text such as code, but also commonly introduce errors and hallucinations in their output. When used to assist software developers, these models may make mistakes that users must go back and fix, or worse, introduce subtle bugs that users may miss entirely. We propose Randomized Utility-driven Synthesis of Uncertain REgions (R-U-SURE), an approach for building uncertainty-aware suggestions based on a decision-theoretic model of goal-conditioned utility, using random samples from a generative model as a proxy for the unobserved possible intents of the end user. Our technique combines minimum-Bayes-risk decoding, dual decomposition, and decision diagrams in order to efficiently produce structured uncertainty summaries, given only sample access to an arbitrary generative model of code and an optional AST parser. We demonstrate R-U-SURE on three developer-assistance tasks, and show that it can be applied different user interaction patterns without retraining the model and leads to more accurate uncertainty estimates than token-probability baselines. We also release our implementation as an open-source library at https://github.com/google-research/r_u_sure.
****
Automatically Auditing Large Language Models via Discrete Optimization
Erik Jones, Anca Dragan, Aditi Raghunathan, Jacob Steinhardt
Auditing large language models for unexpected behaviors is critical to preempt catastrophic deployments, yet remains challenging. In this work, we cast auditing as an optimization problem, where we automatically search for input-output pairs that match a desired target behavior. For example, we might aim to find a non-toxic input that starts with “Barack Obama” that a model maps to a toxic output. This optimization problem is difficult to solve as the set of feasible points is sparse, the space is discrete, and the language models we audit are non-linear and high-dimensional. To combat these challenges, we introduce a discrete optimization algorithm, ARCA, that jointly and efficiently optimizes over inputs and outputs. Our approach automatically uncovers derogatory completions about celebrities (e.g. "Barack Obama is a legalized unborn" –$>$ "child murderer"), produces French inputs that complete to English outputs, and finds inputs that generate a specific name. Our work offers a promising new tool to uncover models’ failure-modes before deployment. Content Warning: This paper contains examples that may be offensive in nature.
****
On the Expressive Power of Geometric Graph Neural Networks

Chaitanya K. Joshi, Cristian Bodnar, Simon V Mathis, Taco Cohen, Pietro Lio
The expressive power of Graph Neural Networks (GNNs) has been studied extensively through the Weisfeiler-Leman (WL) graph isomorphism test. However, standard GNNs and the WL framework are inapplicable for geometric graphs embedded in Euclidean space, such as biomolecules, materials, and other physical systems. In this work, we propose a geometric version of the WL test (GWL) for discriminating geometric graphs while respecting the underlying physical symmetries: permutations, rotation, reflection, and translation. We use GWL to characterise the expressive power of geometric GNNs that are invariant or equivariant to physical symmetries in terms of distinguishing geometric graphs. GWL unpacks how key design choices influence geometric GNN expressivity: (1) Invariant layers have limited expressivity as they cannot distinguish one-hop identical geometric graphs; (2) Equivariant layers distinguish a larger class of graphs by propagating geometric information beyond local neighbourhoods; (3) Higher order tensors and scalarisation enable maximally powerful geometric GNNs; and (4) GWL's discrimination-based perspective is equivalent to universal approximation. Synthetic experiments supplementing our results are available at https://github.com/chaitjo/geometric-gnn-dojo
****
Data-Efficient Contrastive Self-supervised Learning: Most Beneficial Examples for Supervised Learning Contribute the Least
Siddharth Joshi, Baharan Mirzasoleiman
Self-supervised learning (SSL) learns high-quality representations from large pools of unlabeled training data. As datasets grow larger, it becomes crucial to identify the examples that contribute the most to learning such representations. This enables efficient SSL by reducing the volume of data required. Nevertheless, quantifying the value of examples for SSL has remained an open question. In this work, we address this problem for the first time, by proving that examples that contribute the most to contrastive SSL are those that have the most similar augmentations to other examples, in expectation. We provide rigorous guarantees for the generalization performance of contrastive learning on such subsets. Through extensive experiments, we show that we can safely exclude 20% of examples from CIFAR100 and 40% from STL10 and TinyImageNet, without affecting downstream task performance. In general, subsets selected by our method outperform random subsets by over 3% across these datasets. Interestingly, we also discover the subsets that contribute the most to contrastive learning are those that contribute the least to supervised learning.
****
Robust Subtask Learning for Compositional Generalization
Kishor Jothimurugan, Steve Hsu, Osbert Bastani, Rajeev Alur
Compositional reinforcement learning is a promising approach for training policies to perform complex long-horizon tasks. Typically, a high-level task is decomposed into a sequence of subtasks and a separate policy is trained to perform each subtask. In this paper, we focus on the problem of training subtask policies in a way that they can be used to perform any task; here, a task is given by a sequence of subtasks. We aim to maximize the worst-case performance over all tasks as opposed to the average-case performance. We formulate the problem as a two agent zero-sum game in which the adversary picks the sequence of subtasks. We propose two RL algorithms to solve this game: one is an adaptation of existing multi-agent RL algorithms to our setting and the other is an asynchronous version which enables parallel training of subtask policies. We evaluate our approach on two multi-task environments with continuous states and actions and demonstrate that our algorithms outperform state-of-the-art baselines.
****
On Bridging the Gap between Mean Field and Finite Width Deep Random Multilayer Perceptron with Batch Normalization
Amir Joudaki, Hadi Daneshmand, Francis Bach
Mean-field theory is widely used in theoretical studies of neural networks. In this paper, we analyze the role of depth in the concentration of mean-field predictions for Gram matrices of hidden representations in deep multilayer perceptron

(MLP) with batch normalization (BN) at initialization. It is postulated that th
e mean-field predictions suffer from layer-wise errors that amplify with depth.
We demonstrate that BN avoids this error amplification with depth. When the chai
n of hidden representations is rapidly mixing, we establish a concentration boun
d for a mean-field model of Gram matrices. To our knowledge, this is the first c
oncentration bound that does not become vacuous with depth for standard MLPs wit
h a finite width.
****
FARE: Provably Fair Representation Learning with Practical Certificates
Nikola Jovanovi■, Mislav Balunovic, Dimitar Iliev Dimitrov, Martin Vechev
Fair representation learning (FRL) is a popular class of methods aiming to produ
ce fair classifiers via data preprocessing. Recent regulatory directives stress
the need for FRL methods that provide practical certificates, i.e., provable upp
er bounds on the unfairness of any downstream classifier trained on preprocessed
data, which directly provides assurance in a practical scenario. Creating such
FRL methods is an important challenge that remains unsolved. In this work, we ad
dress that challenge and introduce FARE (Fairness with Restricted Encoders), the
first FRL method with practical fairness certificates. FARE is based on our key
insight that restricting the representation space of the encoder enables the de
rivation of practical guarantees, while still permitting favorable accuracy-fair
ness tradeoffs for suitable instantiations, such as one we propose based on fair
trees. To produce a practical certificate, we develop and apply a statistical p
rocedure that computes a finite sample high-confidence upper bound on the unfair
ness of any downstream classifier trained on FARE embeddings. In our comprehensi
ve experimental evaluation, we demonstrate that FARE produces practical certific
ates that are tight and often even comparable with purely empirical results obta
ined by prior methods, which establishes the practical value of our approach.
****
Scaling of Class-wise Training Losses for Post-hoc Calibration
Seungjin Jung, Seungmo Seo, Yonghyun Jeong, Jongwon Choi
The class-wise training losses often diverge as a result of the various levels o
f intra-class and inter-class appearance variation, and we find that the divergi
ng class-wise training losses cause the uncalibrated prediction with its reliabi
lity. To resolve the issue, we propose a new calibration method to synchronize t
he class-wise training losses. We design a new training loss to alleviate the va
riance of class-wise training losses by using multiple class-wise scaling factor
s. Since our framework can compensate the training losses of overfitted classes
with those of under-fitted classes, the integrated training loss is preserved, p
reventing the performance drop even after the model calibration. Furthermore, ou
r method can be easily employed in the post-hoc calibration methods, allowing us
to use the pre-trained model as an initial model and reduce the additional comp
utation for model calibration. We validate the proposed framework by employing i
t in the various post-hoc calibration methods, which generally improves calibrat
ion performance while preserving accuracy, and discover through the investigatio
n that our approach performs well with unbalanced datasets and untuned hyperpara
meters.
****
Fighting Fire with Fire: Contrastive Debiasing without Bias-free Data via Genera
tive Bias-transformation
Yeonsung Jung, Hajin Shim, June Yong Yang, Eunho Yang
Deep neural networks (DNNs), despite their ability to generalize with over-capac
ity networks, often rely heavily on the malignant bias as shortcuts instead of t
ask-related information for discriminative tasks. This can lead to poor performa
nce on real-world inputs, particularly when the majority of the sample is biased
. To address the highly biased issue, recent studies either exploit auxiliary in
formation which is rarely obtainable in practice or sift handful bias-free sampl
es to emphasize them for debiasing. However, these methods are not always guaran
teed to work due to unmet presumptions. In this paper, we propose Contrastive De
biasing via Generative Bias-transformation (CDvG) which is capable of operating
without explicitly exploiting bias labels and bias-free samples. Motivated by ou

r observation that not only discriminative models but also image translation mod
els tend to focus on the malignant bias, CDvG employs an image translation model
 to transform the bias to another mode of bias while preserving task-relevant in
formation. Through contrastive learning, the bias-transformed views are set agai
nst each other to learn bias-invariant representations. Our method shows a bette
r debiasing effect when bias is more malignant as opposed to previous methods, a
nd can also be integrated with the methods that focus on bias-free samples in a
plug-and-play manner for further improvement. Experimental results on diverse da
tasets demonstrate that the proposed method outperforms the state-of-the-art, es
pecially when bias-free samples are extremely scarce or absent.
****

Estimating Joint Treatment Effects by Combining Multiple Experiments
Yonghan Jung, Jin Tian, Elias Bareinboim
Estimating the effects of multi-dimensional treatments (i.e., joint treatment ef
fects) is critical in many data-intensive domains, including genetics and drug e
valuation. The main challenges for studying the joint treatment effects include
the need for large sample sizes to explore different treatment combinations as w
ell as potentially unsafe treatment interactions. In this paper, we develop mach
inery for estimating joint treatment effects by combining data from multiple exp
erimental datasets. In particular, first, we develop new identification conditio
ns for determining whether a joint treatment effect can be computed in terms of
multiple interventional distributions under various scenarios. Further, we devel
op estimators with statistically appealing properties, including consistency and
 robustness to model misspecification and slow convergence. Finally, we perform
simulation studies, which corroborate the effectiveness of the proposed methods.
****

The Catalog Problem: Clustering and Ordering Variable-Sized Sets
Mateusz Maria Jurewicz, Graham W. Taylor, Leon Derczynski
Prediction of a $\textbf{varying number}$ of $\textbf{ordered clusters}$ from se
ts of $\textbf{any cardinality}$ is a challenging task for neural networks, comb
ining elements of set representation, clustering and learning to order. This tas
k arises in many diverse areas, ranging from medical triage and early discharge,
 through machine part management and multi-channel signal analysis for petroleum
 exploration to product catalog structure prediction. This paper focuses on that
 last area, which exemplifies a number of challenges inherent to adaptive ordere
d clustering, referred to further as the eponymous $\textit{Catalog Problem}$. T
hese include learning variable cluster constraints, exhibiting relational reason
ing and managing combinatorial complexity. Despite progress in both neural clust
ering and set-to-sequence methods, no joint, fully differentiable model exists t
o-date. We develop such a modular architecture, referred to further as Neural Or
dered Clusters (NOC), enhance it with a specific mechanism for learning cluster-
level cardinality constraints, and provide a robust comparison of its performanc
e in relation to alternative models. We test our method on three datasets, inclu
ding synthetic catalog structures and PROCAT, a dataset of real-world catalogs c
onsisting of over 1.5M products, achieving state-of-the-art results on a new, mo
re challenging formulation of the underlying problem, which has not been address
ed before. Additionally, we examine the network’s ability to learn higher-order
interactions.
****

Equivariance with Learned Canonicalization Functions
Sékou-Oumar Kaba, Arnab Kumar Mondal, Yan Zhang, Yoshua Bengio, Siamak Ravanbakh
sh
Symmetry-based neural networks often constrain the architecture in order to achi
eve invariance or equivariance to a group of transformations. In this paper, we
propose an alternative that avoids this architectural constraint by learning to
produce canonical representations of the data. These canonicalization functions
can readily be plugged into non-equivariant backbone architectures. We offer exp
licit ways to implement them for some groups of interest. We show that this appr
oach enjoys universality while providing interpretable insights. Our main hypoth
esis, supported by our empirical results, is that learning a small neural networ

k to perform canonicalization is better than using predefined heuristics. Our ex
periments show that learning the canonicalization function is competitive with e
xisting techniques for learning equivariant functions across many tasks, includi
ng image classification, N-body dynamics prediction, point cloud classificatio
n and part segmentation, while being faster across the board.
****
Biases in Evaluation of Molecular Optimization Methods and Bias Reduction Strate
gies
Hiroshi Kajino, Kohei Miyaguchi, Takayuki Osogami
We are interested in an evaluation methodology for molecular optimization. Given
 a sample of molecules and their properties of our interest, we wish not only to
 train a generator of molecules optimized with respect to a target property but
also to evaluate its performance accurately. A common practice is to train a pre
dictor of the target property using the sample and apply it to both training and
 evaluating the generator. However, little is known about its statistical proper
ties, and thus, we are not certain about whether this performance estimate is re
liable or not. We theoretically investigate this evaluation methodology and show
 that it potentially suffers from two biases; one is due to misspecification of
the predictor and the other to reusing the same finite sample for training and e
valuation. We discuss bias reduction methods for each of the biases, and empiric
ally investigate their effectiveness.
****
Statistical Indistinguishability of Learning Algorithms
Alkis Kalavasis, Amin Karbasi, Shay Moran, Grigoris Velegkas
When two different parties use the same learning rule on their own data, how can
 we test whether the distributions of the two outcomes are similar? In this pape
r, we study the similarity of outcomes of learning rules through the lens of the
 Total Variation (TV) distance of distributions. We say that a learning rule is
TV indistinguishable if the expected TV distance between the posterior distribut
ions of its outputs, executed on two training data sets drawn independently from
 the same distribution, is small. We first investigate the learnability of hypot
hesis classes using TV indistinguishable learners. Our main results are informat
ion-theoretic equivalences between TV indistinguishability and existing algorith
mic stability notions such as replicability and approximate differential privacy
. Then, we provide statistical amplification and boosting algorithms for TV indi
stinguishable learners.
****
Identifying Interpretable Subspaces in Image Representations
Neha Kalibhat, Shweta Bhardwaj, C. Bayan Bruss, Hamed Firooz, Maziar Sanjabi, So
heil Feizi
We propose Automatic Feature Explanation using Contrasting Concepts (FALCON), an
 interpretability framework to explain features of image representations. For a
target feature, FALCON captions its highly activating cropped images using a lar
ge captioning dataset (like LAION-400m) and a pre-trained vision-language model
like CLIP. Each word among the captions is scored and ranked leading to a small
number of shared, human-understandable concepts that closely describe the target
 feature. FALCON also applies contrastive interpretation using lowly activating
(counterfactual) images, to eliminate spurious concepts. Although many existing
approaches interpret features independently, we observe in state-of-the-art self
-supervised and supervised models, that less than 20% of the representation spac
e can be explained by individual features. We show that features in larger space
s become more interpretable when studied in groups and can be explained with hig
h-order scoring concepts through FALCON. We discuss how extracted concepts can b
e used to explain and debug failures in downstream tasks. Finally, we present a
technique to transfer concepts from one (explainable) representation space to an
other unseen representation space by learning a simple linear transformation.
****
Nonlinear Causal Discovery with Latent Confounders
David Kaltenpoth, Jilles Vreeken
Causal discovery, the task of discovering the causal graph over a set of observe

d variables $X_1,\ldots,X_m$, is a challenging problem. One of the cornerstone a ssumptions is that of causal sufficiency: that all common causes of all measured variables have been observed. When it does not hold, causal discovery algorithm s making this assumption return networks with many spurious edges. In this paper , we propose a nonlinear causal model involving hidden confounders. We show that it is identifiable from only the observed data and propose an efficient method for recovering this causal model. At the heart of our approach is a variational autoencoder which parametrizes both the causal interactions between observed var iables as well as the influence of the unobserved confounders. Empirically we sh ow that it outperforms other state-of-the-art methods for causal discovery under latent confounding on synthetic and real-world data.
****

Deep Generative Symbolic Regression with Monte-Carlo-Tree-Search
Pierre-Alexandre Kamienny, Guillaume Lample, Sylvain Lamprier, Marco Virgolin
Symbolic regression (SR) is the problem of learning a symbolic expression from n umerical data. Recently, deep neural models trained on procedurally-generated sy nthetic datasets showed competitive performance compared to more classical Genet ic Programming (GP) ones. Unlike their GP counterparts, these neural approaches are trained to generate expressions from datasets given as context. This allows them to produce accurate expressions in a single forward pass at test time. Howe ver, they usually do not benefit from search abilities, which result in low perf ormance compared to GP on out-of-distribution datasets. In this paper, we propos e a novel method which provides the best of both worlds, based on a Monte-Carlo Tree Search procedure using a context-aware neural mutation model, which is init ially pre-trained to learn promising mutations, and further refined from success ful experiences in an online fashion. The approach demonstrates state-of-the-art performance on the well-known SRBench benchmark.
****

One-vs-the-Rest Loss to Focus on Important Samples in Adversarial Training
Sekitoshi Kanai, Shin’Ya Yamaguchi, Masanori Yamada, Hiroshi Takahashi, Kentaro Ohno, Yasutoshi Ida
This paper proposes a new loss function for adversarial training. Since adversar ial training has difficulties, e.g., necessity of high model capacity, focusing on important data points by weighting cross-entropy loss has attracted much atte ntion. However, they are vulnerable to sophisticated attacks, e.g., Auto-Attack. This paper experimentally reveals that the cause of their vulnerability is thei r small margins between logits for the true label and the other labels. Since ne ural networks classify the data points based on the logits, logit margins should be large enough to avoid flipping the largest logit by the attacks. Importance- aware methods do not increase logit margins of important samples but decrease th ose of less-important samples compared with cross-entropy loss. To increase logi t margins of important samples, we propose switching one-vs-the-rest loss (SOVR) , which switches from cross-entropy to one-vs-the-rest loss for important sample s that have small logit margins. We prove that one-vs-the-rest loss increases lo git margins two times larger than the weighted cross-entropy loss for a simple p roblem. We experimentally confirm that SOVR increases logit margins of important samples unlike existing methods and achieves better robustness against Auto-Att ack than importance-aware methods.
****

Large Language Models Struggle to Learn Long-Tail Knowledge
Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, Colin Raffel
The Internet contains a wealth of knowledge—from the birthdays of historical fig ures to tutorials on how to code—all of which may be learned by language models. However, while certain pieces of information are ubiquitous on the web, others appear extremely rarely. In this paper, we study the relationship between the kn owledge memorized by large language models and the information in pre-training d atasets scraped from the web. In particular, we show that a language model’s abi lity to answer a fact-based question relates to how many documents associated wi th that question were seen during pre-training. We identify these relevant docum ents by entity linking pre-training datasets and counting documents that contain

the same entities as a given question-answer pair. Our results demonstrate strong correlational and causal relationships between accuracy and relevant document count for numerous question answering datasets (e.g., TriviaQA), pre-training corpora (e.g., ROOTS), and model sizes (e.g., 176B parameters). Moreover, while larger models are better at learning long-tail knowledge, we estimate that today's models must be scaled by many orders of magnitude to reach competitive QA performance on questions with little support in the pre-training data. Finally, we show that retrieval-augmentation can reduce the dependence on relevant pre-training information, presenting a promising approach for capturing the long-tail.

****

Git-Theta: A Git Extension for Collaborative Development of Machine Learning Models

Nikhil Kandpal, Brian Lester, Mohammed Muqeeth, Anisha Mascarenhas, Monty Evans, Vishal Baskaran, Tenghao Huang, Haokun Liu, Colin Raffel

Currently, most machine learning models are trained by centralized teams and are rarely updated. In contrast, open-source software development involves the iterative development of a shared artifact through distributed collaboration using a version control system. In the interest of enabling collaborative and continual improvement of machine learning models (Raffel, 2023), we introduce Git-Theta, a version control system for machine learning models. Git-Theta is an extension to Git, the most widely used version control software, that allows fine-grained tracking of changes to model parameters alongside code and other artifacts. Unlike existing version control systems that treat a model checkpoint as a blob of data, Git-Theta leverages the structure of checkpoints to support communication-efficient updates, automatic model merges, and meaningful reporting about the difference between two versions of a model. In addition, Git-Theta includes a plug-in system that enables users to easily add support for new functionality. In this paper, we introduce Git-Theta's design and features and include an example use-case of Git-Theta where a pre-trained model is continually adapted and modified. We publicly release Git-Theta in hopes of kickstarting a new era of collaborative model development. https://github.com/r-three/git-theta/

****

A Deep Conjugate Direction Method for Iteratively Solving Linear Systems

Ayano Kaneda, Osman Akar, Jingyu Chen, Victoria Alicia Trevino Kala, David Hyde, Joseph Teran

We present a novel deep learning approach to approximate the solution of large, sparse, symmetric, positive-definite linear systems of equations. Motivated by the conjugate gradients algorithm that iteratively selects search directions for minimizing the matrix norm of the approximation error, we design an approach that utilizes a deep neural network to accelerate convergence via data-driven improvement of the search direction at each iteration. Our method leverages a carefully chosen convolutional network to approximate the action of the inverse of the linear operator up to an arbitrary constant. We demonstrate the efficacy of our approach on spatially discretized Poisson equations, which arise in computational fluid dynamics applications, with millions of degrees of freedom. Unlike state-of-the-art learning approaches, our algorithm is capable of reducing the linear system residual to a given tolerance in a small number of iterations, independent of the problem size. Moreover, our method generalizes effectively to various systems beyond those encountered during training.

****

Leveraging Proxy of Training Data for Test-Time Adaptation

Juwon Kang, Nayeong Kim, Donghyeon Kwon, Jungseul Ok, Suha Kwak

We consider test-time adaptation (TTA), the task of adapting a trained model to an arbitrary test domain using unlabeled input data on-the-fly during testing. A common practice of TTA is to disregard data used in training due to large memory demand and privacy leakage. However, the training data are the only source of supervision. This motivates us to investigate a proper way of using them while minimizing the side effects. To this end, we propose two lightweight yet informative proxies of the training data and a TTA method fully exploiting them. One of the proxies is composed of a small number of images synthesized (hence, less pri

vacy-sensitive) by data condensation which minimizes their domain-specificity to capture a general underlying structure over a wide spectrum of domains. Then, in TTA, they are translated into labeled test data by stylizing them to match styles of unlabeled test samples. This enables virtually supervised test-time training. The other proxy is inter-class relations of training data, which are transferred to target model during TTA. On four public benchmarks, our method outperforms the state-of-the-art ones at remarkably less computation and memory.
****
Beyond Reward: Offline Preference-guided Policy Optimization
Yachen Kang, Diyuan Shi, Jinxin Liu, Li He, Donglin Wang
This study focuses on the topic of offline preference-based reinforcement learning (PbRL), a variant of conventional reinforcement learning that dispenses with the need for online interaction or specification of reward functions. Instead, the agent is provided with fixed offline trajectories and human preferences between pairs of trajectories to extract the dynamics and task information, respectively. Since the dynamics and task information are orthogonal, a naive approach would involve using preference-based reward learning followed by an off-the-shelf offline RL algorithm. However, this requires the separate learning of a scalar reward function, which is assumed to be an information bottleneck of the learning process. To address this issue, we propose the offline preference-guided policy optimization (OPPO) paradigm, which models offline trajectories and preferences in a one-step process, eliminating the need for separately learning a reward function. OPPO achieves this by introducing an offline hindsight information matching objective for optimizing a contextual policy and a preference modeling objective for finding the optimal context. OPPO further integrates a well-performing decision policy by optimizing the two objectives iteratively. Our empirical results demonstrate that OPPO effectively models offline preferences and outperforms prior competing baselines, including offline RL algorithms performed over either true or pseudo reward function specifications. Our code is available on the project website: https://sites.google.com/view/oppo-icml-2023.
****
Poisoning Generative Replay in Continual Learning to Promote Forgetting
Siteng Kang, Zhan Shi, Xinhua Zhang
Generative models have grown into the workhorse of many state-of-the-art machine learning methods. However, their vulnerability under poisoning attacks has been largely understudied. In this work, we investigate this issue in the context of continual learning, where generative replayers are utilized to tackle catastrophic forgetting. By developing a novel customization of dirty-label input-aware backdoors to the online setting, our attacker manages to stealthily promote forgetting while retaining high accuracy at the current task and sustaining strong defenders. Our approach taps into an intriguing property of generative models, namely that they cannot well capture input-dependent triggers. Experiments on four standard datasets corroborate the poisoner’s effectiveness.
****
Node Embedding from Neural Hamiltonian Orbits in Graph Neural Networks
Qiyu Kang, Kai Zhao, Yang Song, Sijie Wang, Wee Peng Tay
In the graph node embedding problem, embedding spaces can vary significantly for different data types, leading to the need for different GNN model types. In this paper, we model the embedding update of a node feature as a Hamiltonian orbit over time. Since the Hamiltonian orbits generalize the exponential maps, this approach allows us to learn the underlying manifold of the graph in training, in contrast to most of the existing literature that assumes a fixed graph embedding manifold with a closed exponential map solution. Our proposed node embedding strategy can automatically learn, without extensive tuning, the underlying geometry of any given graph dataset even if it has diverse geometries. We test Hamiltonian functions of different forms and verify the performance of our approach on two graph node embedding downstream tasks: node classification and link prediction. Numerical experiments demonstrate that our approach adapts better to different types of graph datasets than popular state-of-the-art graph node embedding GNNs. The code is available at https://github.com/zknus/Hamiltonian-GNN.

****
Understanding Gradient Regularization in Deep Learning: Efficient Finite-Difference Computation and Implicit Bias
Ryo Karakida, Tomoumi Takase, Tomohiro Hayase, Kazuki Osawa
Gradient regularization (GR) is a method that penalizes the gradient norm of the training loss during training. While some studies have reported that GR can improve generalization performance, little attention has been paid to it from the algorithmic perspective, that is, the algorithms of GR that efficiently improve the performance. In this study, we first reveal that a specific finite-difference computation, composed of both gradient ascent and descent steps, reduces the computational cost of GR. Next, we show that the finite-difference computation also works better in the sense of generalization performance. We theoretically analyze a solvable model, a diagonal linear network, and clarify that GR has a desirable implicit bias to so-called rich regime and finite-difference computation strengthens this bias. Furthermore, finite-difference GR is closely related to some other algorithms based on iterative ascent and descent steps for exploring flat minima. In particular, we reveal that the flooding method can perform finite-difference GR in an implicit way. Thus, this work broadens our understanding of GR for both practice and theory.
****
Langevin Thompson Sampling with Logarithmic Communication: Bandits and Reinforcement Learning
Amin Karbasi, Nikki Lijing Kuang, Yian Ma, Siddharth Mitra
Thompson sampling (TS) is widely used in sequential decision making due to its ease of use and appealing empirical performance. However, many existing analytical and empirical results for TS rely on restrictive assumptions on reward distributions, such as belonging to conjugate families, which limits their applicability in realistic scenarios. Moreover, sequential decision making problems are often carried out in a batched manner, either due to the inherent nature of the problem or to serve the purpose of reducing communication and computation costs. In this work, we jointly study these problems in two popular settings, namely, stochastic multi-armed bandits (MABs) and infinite-horizon reinforcement learning (RL), where TS is used to learn the unknown reward distributions and transition dynamics, respectively. We propose batched Langevin Thompson Sampling algorithms that leverage MCMC methods to sample from approximate posteriors with only logarithmic communication costs in terms of batches. Our algorithms are computationally efficient and maintain the same order-optimal regret guarantees of $\mathcal{O}(\log T)$ for stochastic MABs, and $\mathcal{O}(\sqrt{T})$ for RL. We complement our theoretical findings with experimental results.
****
On the Relationship Between Explanation and Prediction: A Causal View
Amir-Hossein Karimi, Krikamol Muandet, Simon Kornblith, Bernhard Schölkopf, Been Kim
Being able to provide explanations for a model’s decision has become a central requirement for the development, deployment, and adoption of machine learning models. However, we are yet to understand what explanation methods can and cannot do. How do upstream factors such as data, model prediction, hyperparameters, and random initialization influence downstream explanations? While previous work raised concerns that explanations (E) may have little relationship with the prediction (Y), there is a lack of conclusive study to quantify this relationship. Our work borrows tools from causal inference to systematically assay this relationship. More specifically, we study the relationship between E and Y by measuring the treatment effect when intervening on their causal ancestors, i.e., on hyperparameters and inputs used to generate saliency-based Es or Ys. Our results suggest that the relationships between E and Y is far from ideal. In fact, the gap between ’ideal’ case only increase in higher-performing models — models that are likely to be deployed. Our work is a promising first step towards providing a quantitative measure of the relationship between E and Y, which could also inform the future development of methods for E with a quantitative metric.
****

Cocktail Party Attack: Breaking Aggregation-Based Privacy in Federated Learning Using Independent Component Analysis
Sanjay Kariyappa, Chuan Guo, Kiwan Maeng, Wenjie Xiong, G. Edward Suh, Moinuddin K Qureshi, Hsien-Hsin S. Lee
Federated learning (FL) aims to perform privacy-preserving machine learning on distributed data held by multiple data owners. To this end, FL requires the data owners to perform training locally and share the gradients or weight updates (instead of the private inputs) with the central server, which are then securely aggregated over multiple data owners. Although aggregation by itself does not offer provable privacy protection, prior work suggested that if the batch size is sufficiently large the aggregation may be secure enough. In this paper, we propose the Cocktail Party Attack (CPA) that, contrary to prior belief, is able to recover the private inputs from gradients/weight updates aggregated over as many as 1024 samples. CPA leverages the crucial insight that aggregate gradients from a fully connected (FC) layer is a linear combination of its inputs, which allows us to frame gradient inversion as a blind source separation (BSS) problem. We adapt independent component analysis (ICA)—a classic solution to the BSS problem—to recover private inputs for FC and convolutional networks, and show that CPA significantly outperforms prior gradient inversion attacks, scales to ImageNet-sized inputs, and works on large batch sizes of up to 1024.
****
General Sequential Episodic Memory Model
Arjun Karuvally, Terrence Sejnowski, Hava T Siegelmann
The state-of-the-art memory model is the General Associative Memory Model, a generalization of the classical Hopfield network. Like its ancestor, the general associative memory has a well-defined state-dependant energy surface, and its memories correlate with its fixed points. This is unlike human memories, which are commonly sequential rather than separated fixed points. In this paper, we introduce a class of General Sequential Episodic Memory Models (GSEMM) that, in the adiabatic limit, exhibit a dynamic energy surface, leading to a series of meta-stable states capable of encoding memory sequences. A multiple-timescale architecture enables the dynamic nature of the energy surface with newly introduced asymmetric synapses and signal propagation delays. We demonstrate its dense capacity under polynomial activation functions. GSEMM combines separate memories, short and long sequential episodic memories, under a unified theoretical framework, demonstrating how energy-based memory modeling can provide richer, human-like episodes.
****
Regression with Sensor Data Containing Incomplete Observations
Takayuki Katsuki, Takayuki Osogami
This paper addresses a regression problem in which output label values are the results of sensing the magnitude of a phenomenon. A low value of such labels can mean either that the actual magnitude of the phenomenon was low or that the sensor made an incomplete observation. This leads to a bias toward lower values in labels and the resultant learning because labels may have lower values due to incomplete observations, even if the actual magnitude of the phenomenon was high. Moreover, because an incomplete observation does not provide any tags indicating incompleteness, we cannot eliminate or impute them. To address this issue, we propose a learning algorithm that explicitly models incomplete observations corrupted with an asymmetric noise that always has a negative value. We show that our algorithm is unbiased as if it were learned from uncorrupted data that does not involve incomplete observations. We demonstrate the advantages of our algorithm through numerical experiments.
****
Data Representations’ Study of Latent Image Manifolds
Ilya Kaufman, Omri Azencot
Deep neural networks have been demonstrated to achieve phenomenal success in many domains, and yet their inner mechanisms are not well understood. In this paper, we investigate the curvature of image manifolds, i.e., the manifold deviation from being flat in its principal directions. We find that state-of-the-art train

ed convolutional neural networks for image classification have a characteristic curvature profile along layers: an initial steep increase, followed by a long phase of a plateau, and followed by another increase. In contrast, this behavior does not appear in untrained networks in which the curvature flattens. We also show that the curvature gap between the last two layers has a strong correlation with the generalization capability of the network. Moreover, we find that the intrinsic dimension of latent codes is not necessarily indicative of curvature. Finally, we observe that common regularization methods such as mixup yield flatter representations when compared to other methods. Our experiments show consistent results over a variety of deep learning architectures and multiple data sets.
****

Multi-Modal Classifiers for Open-Vocabulary Object Detection
Prannay Kaul, Weidi Xie, Andrew Zisserman
The goal of this paper is open-vocabulary object detection (OVOD) — building a model that can detect objects beyond the set of categories seen at training, thus enabling the user to specify categories of interest at inference without the need for model retraining. We adopt a standard two- stage object detector architecture, and explore three ways for specifying novel categories: via language descriptions, via image exemplars, or via a combination of the two. We make three contributions: first, we prompt a large language model (LLM) to generate informative language descriptions for object classes, and construct powerful text-based classifiers; second, we employ a visual aggregator on image exemplars that can ingest any number of images as input, forming vision-based classifiers; and third, we provide a simple method to fuse information from language descriptions and image exemplars, yield- ing a multi-modal classifier. When evaluating on the challenging LVIS open-vocabulary bench- mark we demonstrate that: (i) our text-based classifiers outperform all previous OVOD works; (ii) our vision-based classifiers perform as well as text-based classifiers in prior work; (iii) using multi-modal classifiers perform better than either modality alone; and finally, (iv) our text-based and multi-modal classifiers yield better performance than a fully-supervised detector.
****

Learning Mixtures of Markov Chains and MDPs
Chinmaya Kausik, Kevin Tan, Ambuj Tewari
We present an algorithm for learning mixtures of Markov chains and Markov decision processes (MDPs) from short unlabeled trajectories. Specifically, our method handles mixtures of Markov chains with optional control input by going through a multi-step process, involving (1) a subspace estimation step, (2) spectral clustering of trajectories using "pairwise distance estimators," along with refinement using the EM algorithm, (3) a model estimation step, and (4) a classification step for predicting labels of new trajectories. We provide end-to-end performance guarantees, where we only explicitly require the length of trajectories to be linear in the number of states and the number of trajectories to be linear in a mixing time parameter. Experimental results support these guarantees, where we attain 96.6% average accuracy on a mixture of two MDPs in gridworld, outperforming the EM algorithm with random initialization (73.2% average accuracy). We also significantly outperform the EM algorithm on real data from the LastFM song dataset.
****

Curious Replay for Model-based Adaptation
Isaac Kauvar, Chris Doyle, Linqi Zhou, Nick Haber
Agents must be able to adapt quickly as an environment changes. We find that existing model-based reinforcement learning agents are unable to do this well, in part because of how they use past experiences to train their world model. Here, we present Curious Replay—a form of prioritized experience replay tailored to model-based agents through use of a curiosity-based priority signal. Agents using Curious Replay exhibit improved performance in an exploration paradigm inspired by animal behavior and on the Crafter benchmark. DreamerV3 with Curious Replay surpasses state-of-the-art performance on Crafter, achieving a mean score of 19.4 that substantially improves on the previous high score of 14.5 by DreamerV3 with

uniform replay, while also maintaining similar performance on the Deepmind Control Suite. Code for Curious Replay is available at github.com/AutonomousAgentsLab/curiousreplay.

****

How Does Information Bottleneck Help Deep Learning?

Kenji Kawaguchi, Zhun Deng, Xu Ji, Jiaoyang Huang

Numerous deep learning algorithms have been inspired by and understood via the notion of information bottleneck, where unnecessary information is (often implicitly) minimized while task-relevant information is maximized. However, a rigorous argument for justifying why it is desirable to control information bottlenecks has been elusive. In this paper, we provide the first rigorous learning theory for justifying the benefit of information bottleneck in deep learning by mathematically relating information bottleneck to generalization errors. Our theory proves that controlling information bottleneck is one way to control generalization errors in deep learning, although it is not the only or necessary way. We investigate the merit of our new mathematical findings with experiments across a range of architectures and learning settings. In many cases, generalization errors are shown to correlate with the degree of information bottleneck: i.e., the amount of the unnecessary information at hidden layers. This paper provides a theoretical foundation for current and future methods through the lens of information bottleneck. Our new generalization bounds scale with the degree of information bottleneck, unlike the previous bounds that scale with the number of parameters, VC dimension, Rademacher complexity, stability or robustness. Our code is publicly available at: https://github.com/xu-ji/information-bottleneck

****

Instrumental Variable Estimation of Average Partial Causal Effects

Yuta Kawakami, Manabu Kuroki, Jin Tian

Instrumental variable (IV) analysis is a powerful tool widely used to elucidate causal relationships. We study the problem of estimating the average partial causal effect (APCE) of a continuous treatment in an IV setting. Specifically, we develop new methods for estimating APCE based on a recent identification condition via an integral equation. We develop two families of methods, nonparametric and parametric - the former uses the Picard iteration to solve the integral equation; the latter parameterizes APCE using a linear basis function model. We analyze the statistical and computational properties of the proposed methods and illustrate them on synthetic and real data.

****

The Test of Tests: A Framework for Differentially Private Hypothesis Testing

Zeki Kazan, Kaiyan Shi, Adam Groce, Andrew P Bray

We present a generic framework for creating differentially private versions of any hypothesis test in a black-box way. We analyze the resulting tests analytically and experimentally. Most crucially, we show good practical performance for small data sets, showing that at $\varepsilon = 1$ we only need 5-6 times as much data as in the fully public setting. We compare our work to the one existing framework of this type, as well as to several individually-designed private hypothesis tests. Our framework is higher power than other generic solutions and at least competitive with (and often better than) individually-designed tests.

****

Exact Inference in High-order Structured Prediction

Chuyang Ke, Jean Honorio

In this paper, we study the problem of inference in high-order structured prediction tasks. In the context of Markov random fields, the goal of a high-order inference task is to maximize a score function on the space of labels, and the score function can be decomposed into sum of unary and high-order potentials. We apply a generative model approach to study the problem of high-order inference, and provide a two-stage convex optimization algorithm for exact label recovery. We also provide a new class of hypergraph structural properties related to hyperedge expansion that drives the success in general high-order inference problems. Finally, we connect the performance of our algorithm and the hyperedge expansion property using a novel hypergraph Cheeger-type inequality.

****
Neural Wave Machines: Learning Spatiotemporally Structured Representations with Locally Coupled Oscillatory Recurrent Neural Networks
T. Anderson Keller, Max Welling
Traveling waves have been measured at a diversity of regions and scales in the brain, however a consensus as to their computational purpose has yet to be reached. An intriguing hypothesis is that traveling waves serve to structure neural representations both in space and time, thereby acting as an inductive bias towards natural data. In this work, we investigate this hypothesis by introducing the Neural Wave Machine (NWM) – a locally coupled oscillatory recurrent neural network capable of exhibiting traveling waves in its hidden state. After training on simple dynamic sequences, we show that this model indeed learns static spatial structure such as topographic organization, and further uses complex spatiotemporal structure such as traveling waves to encode observed transformations. To measure the computational implications of this structure, we use a suite of sequence classification and physical dynamics modeling tasks to show that the NWM is both more parameter efficient, and is able to forecast future trajectories of simple physical dynamical systems more accurately than existing state of the art counterparts.
****
Homomorphism AutoEncoder -- Learning Group Structured Representations from Observed Transitions
Hamza Keurti, Hsiao-Ru Pan, Michel Besserve, Benjamin F Grewe, Bernhard Schölkopf
How can agents learn internal models that veridically represent interactions with the real world is a largely open question. As machine learning is moving towards representations containing not just observational but also interventional knowledge, we study this problem using tools from representation learning and group theory. We propose methods enabling an agent acting upon the world to learn internal representations of sensory information that are consistent with actions that modify it. We use an autoencoder equipped with a group representation acting on its latent space, trained using an equivariance-derived loss in order to enforce a suitable homomorphism property on the group representation. In contrast to existing work, our approach does not require prior knowledge of the group and does not restrict the set of actions the agent can perform. We motivate our method theoretically, and show empirically that it can learn a group representation of the actions, thereby capturing the structure of the set of transformations applied to the environment. We further show that this allows agents to predict the effect of sequences of future actions with improved accuracy.
****
Rethinking Backdoor Attacks
Alaa Khaddaj, Guillaume Leclerc, Aleksandar Makelov, Kristian Georgiev, Hadi Salman, Andrew Ilyas, Aleksander Madry
In a backdoor attack, an adversary inserts maliciously constructed backdoor examples into a training set to make the resulting model vulnerable to manipulation. Defending against such attacks involves viewing inserted examples as outliers in the training set and using techniques from robust statistics to detect and remove them. In this work, we present a different approach to the backdoor attack problem. Specifically, we show that without structural information about the training data distribution, backdoor attacks are indistinguishable from naturally-occuring features in the data—and thus impossible to "detect" in a general sense. Then, guided by this observation, we revisit existing defenses against backdoor attacks and characterize the (often latent) assumptions they make, and on which they depend. Finally, we explore an alternative perspective on backdoor attacks: one that assumes these attacks correspond to the strongest feature in the training data. Under this assumption (which we make formal) we develop a new primitive for detecting backdoor attacks. Our primitive naturally gives rise to a detection algorithm that comes with theoretical guarantees, and is effective in practice.
****

PAC Prediction Sets for Large Language Models of Code
Adam Khakhar, Stephen Mell, Osbert Bastani
Prediction sets have recently been shown to be a promising strategy for quantifying the uncertainty of deep neural networks in a way that provides theoretical guarantees. However, existing techniques have largely targeted settings where the space of labels is simple, so prediction sets can be arbitrary subsets of labels. For structured prediction problems where the space of labels is exponential in size, even prediction sets containing a small fraction of all labels can be exponentially large. In the context of code generation, we propose a solution that considers a restricted set of prediction sets that can compactly be represented as partial programs, which are programs with portions replaced with holes. Given a trained code generation model, our algorithm leverages a programming language's abstract syntax tree to generate a set of programs such that the correct program is in the set with high-confidence. Valuable applications of our algorithm include a Codex-style code generator with holes in uncertain parts of the generated code, which provides a partial program with theoretical guarantees. We evaluate our approach on PICARD (a T5 model for SQL semantic parsing) and Codex (a GPT model for over a dozen programming languages, including Python), demonstrating that our approach generates compact PAC prediction sets. This is the first research contribution that generates PAC prediction sets for generative code models.
****
Accelerated Primal-Dual Methods for Convex-Strongly-Concave Saddle Point Problems
Mohammad Khalafi, Digvijay Boob
We investigate a primal-dual (PD) method for the saddle point problem (SPP) that uses a linear approximation of the primal function instead of the standard proximal step, resulting in a linearized PD (LPD) method. For convex-strongly concave SPP, we observe that the LPD method has a suboptimal dependence on the Lipschitz constant of the primal function. To fix this issue, we combine features of Accelerated Gradient Descent with the LPD method resulting in a single-loop Accelerated Linearized Primal-Dual (ALPD) method. ALPD method achieves the optimal gradient complexity when the SPP has a semi-linear coupling function. We also present an inexact ALPD method for SPPs with a general nonlinear coupling function that maintains the optimal gradient evaluations of the primal parts and significantly improves the gradient evaluations of the coupling term compared to the ALPD method. We verify our findings with numerical experiments.
****
Loss Balancing for Fair Supervised Learning
Mohammad Mahdi Khalili, Xueru Zhang, Mahed Abroshan
Supervised learning models have been used in various domains such as lending, college admission, face recognition, natural language processing, etc. However, they may inherit pre-existing biases from training data and exhibit discrimination against protected social groups. Various fairness notions have been proposed to address unfairness issues. In this work, we focus on Equalized Loss (EL), a fairness notion that requires the expected loss to be (approximately) equalized across different groups. Imposing EL on the learning process leads to a non-convex optimization problem even if the loss function is convex, and the existing fair learning algorithms cannot properly be adopted to find the fair predictor under the EL constraint. This paper introduces an algorithm that can leverage off-the-shelf convex programming tools (e.g., CVXPY (Diamond and Boyd, 2016; Agrawal et al., 2018)) to efficiently find the global optimum of this non-convex optimization. In particular, we propose the ELminimizer algorithm, which finds the optimal fair predictor under EL by reducing the non-convex optimization to a sequence of convex optimization problems. We theoretically prove that our algorithm finds the global optimal solution under certain conditions. Then, we support our theoretical results through several empirical studies
****
Linearly Constrained Bilevel Optimization: A Smoothed Implicit Gradient Approach
Prashant Khanduri, Ioannis Tsaknakis, Yihua Zhang, Jia Liu, Sijia Liu, Jiawei Zhang, Mingyi Hong

This work develops analysis and algorithms for solving a class of bilevel optimi
zation problems where the lower-level (LL) problems have linear constraints. Mos
t of the existing approaches for constrained bilevel problems rely on value func
tion-based approximate reformulations, which suffer from issues such as non-conv
ex and non-differentiable constraints. In contrast, in this work, we develop an
implicit gradient-based approach, which is easy to implement, and is suitable fo
r machine learning applications. We first provide an in-depth understanding of t
he problem, by showing that the implicit objective for such problems is in gener
al non-differentiable. However, if we add some small (linear) perturbation to th
e LL objective, the resulting implicit objective becomes differentiable almost s
urely. This key observation opens the door for developing (deterministic and sto
chastic) gradient-based algorithms similar to the state-of-the-art ones for unco
nstrained bi-level problems. We show that when the implicit function is assumed
to be strongly-convex, convex, and weakly-convex, the resulting algorithms conve
rge with guaranteed rate. Finally, we experimentally corroborate the theoretical
findings and evaluate the performance of the proposed framework on numerical an
d adversarial learning problems.
****

Emergent Asymmetry of Precision and Recall for Measuring Fidelity and Diversity of Generative Models in High Dimensions
Mahyar Khayatkhoei, Wael Abdalmageed
Precision and Recall are two prominent metrics of generative performance, which
were proposed to separately measure the fidelity and diversity of generative mod
els. Given their central role in comparing and improving generative models, unde
rstanding their limitations are crucially important. To that end, in this work,
we identify a critical flaw in the common approximation of these metrics using k
-nearest-neighbors, namely, that the very interpretations of fidelity and divers
ity that are assigned to Precision and Recall can fail in high dimensions, resul
ting in very misleading conclusions. Specifically, we empirically and theoretica
lly show that as the number of dimensions grows, two model distributions with su
pports at equal point-wise distance from the support of the real distribution, c
an have vastly different Precision and Recall regardless of their respective dis
tributions, hence an emergent asymmetry in high dimensions. Based on our theoret
ical insights, we then provide simple yet effective modifications to these metri
cs to construct symmetric metrics regardless of the number of dimensions. Finall
y, we provide experiments on real-world datasets to illustrate that the identifi
ed flaw is not merely a pathological case, and that our proposed metrics are eff
ective in alleviating its impact.
****

Learning-augmented private algorithms for multiple quantile release
Mikhail Khodak, Kareem Amin, Travis Dick, Sergei Vassilvitskii
When applying differential privacy to sensitive data, we can often improve perfo
rmance using external information such as other sensitive data, public data, or
human priors. We propose to use the learning-augmented algorithms (or algorithms
with predictions) framework—previously applied largely to improve time complexi
ty or competitive ratios—as a powerful way of designing and analyzing privacy-pr
eserving methods that can take advantage of such external information to improve
utility. This idea is instantiated on the important task of multiple quantile r
elease, for which we derive error guarantees that scale with a natural measure o
f prediction quality while (almost) recovering state-of-the-art prediction-indep
endent guarantees. Our analysis enjoys several advantages, including minimal ass
umptions about the data, a natural way of adding robustness, and the provision o
f useful surrogate losses for two novel ”meta” algorithms that learn predictions
from other (potentially sensitive) data. We conclude with experiments on challe
nging tasks demonstrating that learning predictions across one or more instances
can lead to large error reductions while preserving privacy.
****

CrossSplit: Mitigating Label Noise Memorization through Data Splitting
Jihye Kim, Aristide Baratin, Yan Zhang, Simon Lacoste-Julien
We approach the problem of improving robustness of deep learning algorithms in t

he presence of label noise. Building upon existing label correction and co-teach ing methods, we propose a novel training procedure to mitigate the memorization of noisy labels, called CrossSplit, which uses a pair of neural networks trained on two disjoint parts of the labeled dataset. CrossSplit combines two main ingr edients: (i) Cross-split label correction. The idea is that, since the model tra ined on one part of the data cannot memorize example-label pairs from the other part, the training labels presented to each network can be smoothly adjusted by using the predictions of its peer network; (ii) Cross-split semi-supervised trai ning. A network trained on one part of the data also uses the unlabeled inputs o f the other part. Extensive experiments on CIFAR-10, CIFAR-100, Tiny-ImageNet an d mini-WebVision datasets demonstrate that our method can outperform the current state-of-the-art in a wide range of noise ratios. The project page is at https: //rlawlgul.github.io/.

****

Trainability, Expressivity and Interpretability in Gated Neural ODEs

Timothy Doyeon Kim, Tankut Can, Kamesh Krishnamurthy

Understanding how the dynamics in biological and artificial neural networks impl ement the computations required for a task is a salient open question in machine learning and neuroscience. In particular, computations requiring complex memory storage and retrieval pose a significant challenge for these networks to implem ent or learn. Recently, a family of models described by neural ordinary differen tial equations (nODEs) has emerged as powerful dynamical neural network models c apable of capturing complex dynamics. Here, we extend nODEs by endowing them wit h adaptive timescales using gating interactions. We refer to these as gated neur al ODEs (gnODEs). Using a task that requires memory of continuous quantities, we demonstrate the inductive bias of the gnODEs to learn (approximate) continuous attractors. We further show how reduced-dimensional gnODEs retain their modeling power while greatly improving interpretability, even allowing explicit visualiz ation of the structure of learned attractors. We introduce a novel measure of ex pressivity which probes the capacity of a neural network to generate complex tra jectories. Using this measure, we explore how the phase-space dimension of the n ODEs and the complexity of the function modeling the flow field contribute to ex pressivity. We see that a more complex function for modeling the flow field allo ws a lower-dimensional nODE to capture a given target dynamics. Finally, we demo nstrate the benefit of gating in nODEs on several real-world tasks.

****

SAAL: Sharpness-Aware Active Learning

Yoon-Yeong Kim, Youngjae Cho, Joonho Jang, Byeonghu Na, Yeongmin Kim, Kyungwoo S ong, Wanmo Kang, Il-Chul Moon

While deep neural networks play significant roles in many research areas, they a re also prone to overfitting problems under limited data instances. To overcome overfitting, this paper introduces the first active learning method to incorpora te the sharpness of loss space into the acquisition function. Specifically, our proposed method, Sharpness-Aware Active Learning (SAAL), constructs its acquisit ion function by selecting unlabeled instances whose perturbed loss becomes maxim um. Unlike the Sharpness-Aware learning with fully-labeled datasets, we design a pseudo-labeling mechanism to anticipate the perturbed loss w.r.t. the ground-tr uth label, which we provide the theoretical bound for the optimization. We condu ct experiments on various benchmark datasets for vision-based tasks in image cla ssification, object detection, and domain adaptive semantic segmentation. The ex perimental results confirm that SAAL outperforms the baselines by selecting inst ances that have the potentially maximal perturbation on the loss. The code is av ailable at https://github.com/YoonyeongKim/SAAL.

****

Demonstration-free Autonomous Reinforcement Learning via Implicit and Bidirectio nal Curriculum

Jigang Kim, Daesol Cho, H. Jin Kim

While reinforcement learning (RL) has achieved great success in acquiring comple x skills solely from environmental interactions, it assumes that resets to the i nitial state are readily available at the end of each episode. Such an assumptio

n hinders the autonomous learning of embodied agents due to the time-consuming a nd cumbersome workarounds for resetting in the physical world. Hence, there has been a growing interest in autonomous RL (ARL) methods that are capable of learn ing from non-episodic interactions. However, existing works on ARL are limited b y their reliance on prior data and are unable to learn in environments where tas k-relevant interactions are sparse. In contrast, we propose a demonstration-free ARL algorithm via Implicit and Bi-directional Curriculum (IBC). With an auxilia ry agent that is conditionally activated upon learning progress and a bidirectio nal goal curriculum based on optimal transport, our method outperforms previous methods, even the ones that leverage demonstrations.

****

Improved Algorithms for Multi-period Multi-class Packing Problems with Bandit Fe edback

Wonyoung Kim, Garud Iyengar, Assaf Zeevi

We consider the linear contextual multi-class multi-period packing problem (LMMP) where the goal is to pack items such that the total vector of consumption is b elow a given budget vector and the total value is as large as possible. We consi der the setting where the reward and the consumption vector associated with each action is a class-dependent linear function of the context, and the decision-ma ker receives bandit feedback. LMMP includes linear contextual bandits with knaps acks and online revenue management as special cases. We establish a new estimato r which guarantees a faster convergence rate, and consequently, a lower regret i n LMMP. We propose a bandit policy that is a closed-form function of said estima ted parameters. When the contexts are non-degenerate, the regret of the proposed policy is sublinear in the context dimension, the number of classes, and the ti me horizon T when the budget grows at least as $\sqrt{T}$. We also resolve an open problem posed in Agrawal & Devanur (2016) and extend the result to a multi-class setting. Our numerical experiments clearly demonstrate that the performanc e of our policy is superior to other benchmarks in the literature.

****

Efficient Latency-Aware CNN Depth Compression via Two-Stage Dynamic Programming

Jinuk Kim, Yeonwoo Jeong, Deokjae Lee, Hyun Oh Song

Recent works on neural network pruning advocate that reducing the depth of the n etwork is more effective in reducing run-time memory usage and accelerating infe rence latency than reducing the width of the network through channel pruning. In this regard, some recent works propose depth compression algorithms that merge convolution layers. However, the existing algorithms have a constricted search s pace and rely on human-engineered heuristics. In this paper, we propose a novel depth compression algorithm which targets general convolution operations. We pro pose a subset selection problem that replaces inefficient activation layers with identity functions and optimally merges consecutive convolution operations into shallow equivalent convolution operations for efficient end-to-end inference la tency. Since the proposed subset selection problem is NP-hard, we formulate a su rrogate optimization problem that can be solved exactly via two-stage dynamic pr ogramming within a few seconds. We evaluate our methods and baselines by TensorR T for a fair inference latency comparison. Our method outperforms the baseline m ethod with higher accuracy and faster inference speed in MobileNetV2 on the Imag eNet dataset. Specifically, we achieve $1.41\times$ speed-up with 0.11%p accur acy gain in MobileNetV2-1.0 on the ImageNet.

****

Probabilistic Concept Bottleneck Models

Eunji Kim, Dahuin Jung, Sangha Park, Siwon Kim, Sungroh Yoon

Interpretable models are designed to make decisions in a human-interpretable man ner. Representatively, Concept Bottleneck Models (CBM) follow a two-step process of concept prediction and class prediction based on the predicted concepts. CBM provides explanations with high-level concepts derived from concept predictions ; thus, reliable concept predictions are important for trustworthiness. In this study, we address the ambiguity issue that can harm reliability. While the exist ence of a concept can often be ambiguous in the data, CBM predicts concepts dete rministically without considering this ambiguity. To provide a reliable interpre

tation against this ambiguity, we propose Probabilistic Concept Bottleneck Models (ProbCBM). By leveraging probabilistic concept embeddings, ProbCBM models uncertainty in concept prediction and provides explanations based on the concept and its corresponding uncertainty. This uncertainty enhances the reliability of the explanations. Furthermore, as class uncertainty is derived from concept uncertainty in ProbCBM, we can explain class uncertainty by means of concept uncertainty. Code is publicly available at https://github.com/ejkim47/prob-cbm.
****

DevFormer: A Symmetric Transformer for Context-Aware Device Placement
Haeyeon Kim, Minsu Kim, Federico Berto, Joungho Kim, Jinkyoo Park
In this paper, we present DevFormer, a novel transformer-based architecture for addressing the complex and computationally demanding problem of hardware design optimization. Despite the demonstrated efficacy of transformers in domains including natural language processing and computer vision, their use in hardware design has been limited by the scarcity of offline data. Our approach addresses this limitation by introducing strong inductive biases such as relative positional embeddings and action-permutation symmetricity that effectively capture the hardware context and enable efficient design optimization with limited offline data. We apply DevFormer to the problem of decoupling capacitor placement and show that it outperforms state-of-the-art methods in both simulated and real hardware, leading to improved performances while reducing the number of components by more than 30%. Finally, we show that our approach achieves promising results in other offline contextual learning-based combinatorial optimization tasks.
****

Refining Generative Process with Discriminator Guidance in Score-based Diffusion Models
Dongjun Kim, Yeongmin Kim, Se Jung Kwon, Wanmo Kang, Il-Chul Moon
The proposed method, Discriminator Guidance, aims to improve sample generation of pre-trained diffusion models. The approach introduces a discriminator that gives explicit supervision to a denoising sample path whether it is realistic or not. Unlike GANs, our approach does not require joint training of score and discriminator networks. Instead, we train the discriminator after score training, making discriminator training stable and fast to converge. In sample generation, we add an auxiliary term to the pre-trained score to deceive the discriminator. This term corrects the model score to the data score at the optimal discriminator, which implies that the discriminator helps better score estimation in a complementary way. Using our algorithm, we achive state-of-the-art results on ImageNet 256x256 with FID 1.83 and recall 0.64, similar to the validation data’s FID (1.68) and recall (0.66). We release the code at https://github.com/alsdudrla10/DG.
****

Robust Non-Linear Feedback Coding via Power-Constrained Deep Learning
Junghoon Kim, Taejoon Kim, David Love, Christopher Brinton
The design of codes for feedback-enabled communications has been a long-standing open problem. Recent research on non-linear, deep learning-based coding schemes have demonstrated significant improvements in communication reliability over linear codes, but are still vulnerable to the presence of forward and feedback noise over the channel. In this paper, we develop a new family of non-linear feedback codes that greatly enhance robustness to channel noise. Our autoencoder-based architecture is designed to learn codes based on consecutive blocks of bits, which obtains de-noising advantages over bit-by-bit processing to help overcome the physical separation between the encoder and decoder over a noisy channel. Moreover, we develop a power control layer at the encoder to explicitly incorporate hardware constraints into the learning optimization, and prove that the resulting average power constraint is satisfied asymptotically. Numerical experiments demonstrate that our scheme outperforms state-of-the-art feedback codes by wide margins over practical forward and feedback noise regimes, and provide information-theoretic insights on the behavior of our non-linear codes. Moreover, we observe that, in a long blocklength regime, canonical error correction codes are still preferable to feedback codes when the feedback noise becomes high. Our code is available at https://anonymous.4open.science/r/RCode1.

****
LESSON: Learning to Integrate Exploration Strategies for Reinforcement Learning via an Option Framework
Woojun Kim, Jeonghye Kim, Youngchul Sung
In this paper, a unified framework for exploration in reinforcement learning (RL) is proposed based on an option-critic architecture. The proposed framework learns to integrate a set of diverse exploration strategies so that the agent can adaptively select the most effective exploration strategy to realize an effective exploration-exploitation trade-off for each given task. The effectiveness of the proposed exploration framework is demonstrated by various experiments in the MiniGrid and Atari environments.
****
BPipe: Memory-Balanced Pipeline Parallelism for Training Large Language Models
Taebum Kim, Hyoungjoo Kim, Gyeong-In Yu, Byung-Gon Chun
Pipeline parallelism is a key technique for training large language models within GPU clusters. However, it often leads to a memory imbalance problem, where certain GPUs face high memory pressure while others underutilize their capacity. This imbalance results in suboptimal training performance, even when the overall GPU memory capacity is sufficient for more efficient setups. To address this inefficiency, we propose BPipe, a novel approach for achieving memory balance in pipeline parallelism. BPipe employs an activation balancing method to transfer intermediate activations between GPUs during training, enabling all GPUs to utilize comparable amounts of memory. With balanced memory utilization, BPipe enhances the training efficiency of large language models like GPT-3 by eliminating redundant recomputations or increasing the micro-batch size. Our evaluation conducted on 48 A100 GPUs across six nodes interconnected with HDR InfiniBand shows that BPipe accelerates the training of GPT-3 96B and GPT-3 134B models by 1.25x-2.17x compared to Megatron-LM, a state-of-the-art framework for training large language models.
****
Probabilistic Imputation for Time-series Classification with Missing Data
Seunghyun Kim, Hyunsu Kim, Eunggu Yun, Hwangrae Lee, Jaehun Lee, Juho Lee
Multivariate time series data for real-world applications typically contain a significant amount of missing values. The dominant approach for classification with such missing values is to impute them heuristically with specific values (zero, mean, values of adjacent time-steps) or learnable parameters. However, these simple strategies do not take the data generative process into account, and more importantly, do not effectively capture the uncertainty in prediction due to the multiple possibilities for the missing values. In this paper, we propose a novel probabilistic framework for classification with multivariate time series data with missing values. Our model consists of two parts; a deep generative model for missing value imputation and a classifier. Extending the existing deep generative models to better capture structures of time-series data, our deep generative model part is trained to impute the missing values in multiple plausible ways, effectively modeling the uncertainty of the imputation. The classifier part takes the time series data along with the imputed missing values and classifies signals, and is trained to capture the predictive uncertainty due to the multiple possibilities of imputations. Importantly, we show that naïvely combining the generative model and the classifier could result in trivial solutions where the generative model does not produce meaningful imputations. To resolve this, we present a novel regularization technique that can promote the model to produce useful imputation values that help classification. Through extensive experiments on real-world time series data with missing values, we demonstrate the effectiveness of our method.
****
Variational Curriculum Reinforcement Learning for Unsupervised Discovery of Skills
Seongun Kim, Kyowoon Lee, Jaesik Choi
Mutual information-based reinforcement learning (RL) has been proposed as a promising framework for retrieving complex skills autonomously without a task-orient

ed reward function through mutual information (MI) maximization or variational e mpowerment. However, learning complex skills is still challenging, due to the fa ct that the order of training skills can largely affect sample efficiency. Inspi red by this, we recast variational empowerment as curriculum learning in goal-co nditioned RL with an intrinsic reward function, which we name Variational Curric ulum RL (VCRL). From this perspective, we propose a novel approach to unsupervis ed skill discovery based on information theory, called Value Uncertainty Variati onal Curriculum (VUVC). We prove that, under regularity conditions, VUVC acceler ates the increase of entropy in the visited states compared to the uniform curri culum. We validate the effectiveness of our approach on complex navigation and r obotic manipulation tasks in terms of sample efficiency and state coverage speed . We also demonstrate that the skills discovered by our method successfully comp lete a real-world robot navigation task in a zero-shot setup and that incorporat ing these skills with a global planner further increases the performance.
****

Margin-based Neural Network Watermarking
Byungjoo Kim, Suyoung Lee, Seanie Lee, Sooel Son, Sung Ju Hwang
As Machine Learning as a Service (MLaaS) platforms become prevalent, deep neural network (DNN) watermarking techniques are gaining increasing attention, which e nables one to verify the ownership of a target DNN model in a black-box scenario . Unfortunately, previous watermarking methods are vulnerable to functionality s tealing attacks, thus allowing an adversary to falsely claim the ownership of a DNN model stolen from its original owner. In this work, we propose a novel margi n-based DNN watermarking approach that is robust to the functionality stealing a ttacks based on model extraction and distillation. Specifically, during training , our method maximizes the margins of watermarked samples by using projected gra dient ascent on them so that their predicted labels cannot change without compro mising the accuracy of the model that the attacker tries to steal. We validate o ur method on multiple benchmarks and show that our watermarking method successfu lly defends against model extraction attacks, outperforming recent baselines.
****

Regularizing Towards Soft Equivariance Under Mixed Symmetries
Hyunsu Kim, Hyungi Lee, Hongseok Yang, Juho Lee
Datasets often have their intrinsic symmetries, and particular deep-learning mod els called equivariant or invariant models have been developed to exploit these symmetries. However, if some or all of these symmetries are only approximate, wh ich frequently happens in practice, these models may be suboptimal due to the ar chitectural restrictions imposed on them. We tackle this issue of approximate sy mmetries in a setup where symmetries are mixed, i.e., they are symmetries of not single but multiple different types and the degree of approximation varies acro ss these types. Instead of proposing a new architectural restriction as in most of the previous approaches, we present a regularizer-based method for building a model for a dataset with mixed approximate symmetries. The key component of our method is what we call equivariance regularizer for a given type of symmetries, which measures how much a model is equivariant with respect to the symmetries o f the type. Our method is trained with these regularizers, one per each symmetry type, and the strength of the regularizers is automatically tuned during traini ng, leading to the discovery of the approximation levels of some candidate symme try types without explicit supervision. Using synthetic function approximation a nd motion forecasting tasks, we demonstrate that our method achieves better accu racy than prior approaches while discovering the approximate symmetry levels cor rectly.
****

Model-based Offline Reinforcement Learning with Count-based Conservatism
Byeongchan Kim, Min-Hwan Oh
In this paper, we present a model-based offline reinforcement learning method th at integrates count-based conservatism, named $\texttt{Count-MORL}$. Our method utilizes the count estimates of state-action pairs to quantify model estimation error, marking the first algorithm of demonstrating the efficacy of count-based conservatism in model-based offline deep RL to the best of our knowledge. For ou

r proposed method, we first show that the estimation error is inversely proporti onal to the frequency of state-action pairs. Secondly, we demonstrate that the l earned policy under the count-based conservative model offers near-optimality pe rformance guarantees. Through extensive numerical experiments, we validate that $\texttt{Count-MORL}$ with hash code implementation significantly outperforms ex isting offline RL algorithms on the D4RL benchmark datasets. The code is accessi ble at https://github.com/oh-lab/Count-MORL.
****
Transformer-based Stagewise Decomposition for Large-Scale Multistage Stochastic Optimization
Chanyeong Kim, Jongwoong Park, Hyunglip Bae, Woo Chang Kim
Solving large-scale multistage stochastic programming (MSP) problems poses a sig nificant challenge as commonly used stagewise decomposition algorithms, includin g stochastic dual dynamic programming (SDDP), face growing time complexity as th e subproblem size and problem count increase. Traditional approaches approximate the value functions as piecewise linear convex functions by incrementally accum ulating subgradient cutting planes from the primal and dual solutions of stagewi se subproblems. Recognizing these limitations, we introduce TranSDDP, a novel Tr ansformer-based stagewise decomposition algorithm. This innovative approach leve rages the structural advantages of the Transformer model, implementing a sequent ial method for integrating subgradient cutting planes to approximate the value f unction. Through our numerical experiments, we affirm TranSDDP’s effectiveness i n addressing MSP problems. It efficiently generates a piecewise linear approxima tion for the value function, significantly reducing computation time while prese rving solution quality, thus marking a promising progression in the treatment of large-scale multistage stochastic programming problems.
****
SurProGenes: Survival Risk-Ordered Representation of Cancer Patients and Genes f or the Identification of Prognostic Genes
Junetae Kim, Kyoungsuk Park, Hanseok Jeong, Youngwook Kim, Jeongseon Kim, Sun-Yo ung Kim
Identifying prognostic genes associated with patient survival is an important go al in cancer genomics, as this information could inform treatment approaches and improve patient outcomes. However, the identification of prognostic genes is co mplicated by the high dimensionality of genetic data, which makes their identifi cation computationally intensive. Furthermore, most cancer genomics studies lack appropriate low-risk groups against which to compare. To address these issues, we present a framework that identifies candidate prognostic genes by integrating representation learning and statistical analysis approaches. Specifically, we p ropose a collaborative filtering-derived mechanism to represent patients in orde r of their survival risk, facilitating their dichotomization. We also propose a mechanism that allows embedded gene vectors to be polarized on the extremities o f, or centered on, both reference axes to facilitate recommendations. Restrictin g our analysis to a few representative genes within each cluster allowed for the efficient identification of prognostic genes. Finally, we demonstrate the poten tial of this proposed framework for identifying prognostic genes.
****
Stable and Consistent Prediction of 3D Characteristic Orientation via Invariant Residual Learning
Seungwook Kim, Chunghyun Park, Yoonwoo Jeong, Jaesik Park, Minsu Cho
Learning to predict reliable characteristic orientations of 3D point clouds is a n important yet challenging problem, as different point clouds of the same class may have largely varying appearances. In this work, we introduce a novel method to decouple the shape geometry and semantics of the input point cloud to achiev e both stability and consistency. The proposed method integrates shape-geometry-based SO(3)-equivariant learning and shape-semantics-based SO(3)-invariant resid ual learning, where a final characteristic orientation is obtained by calibratin g an SO(3)-equivariant orientation hypothesis using an SO(3)-invariant residual rotation. In experiments, the proposed method not only demonstrates superior sta bility and consistency but also exhibits state-of-the-art performances when appl

ied to point cloud part segmentation, given randomly rotated inputs.
****
Prefer to Classify: Improving Text Classifiers via Auxiliary Preference Learning
Jaehyung Kim, Jinwoo Shin, Dongyeop Kang
The development of largely human-annotated benchmarks has driven the success of deep neural networks in various NLP tasks. To enhance the effectiveness of existing benchmarks, collecting new additional input-output pairs is often too costly and challenging, particularly considering their marginal impact on improving the current model accuracy. Instead, additional or complementary annotations on the existing input texts in the benchmarks can be preferable as an efficient way to pay the additional human cost. In this paper, we investigate task-specific preferences between pairs of input texts as a new alternative way for such auxiliary data annotation. From pair-wise comparisons with respect to the task, the auxiliary preference learning enables the model to learn an additional informative training signal that cannot be captured with instance-wise task labels. To this end, we propose a novel multi-task learning framework, called prefer-to-classify (P2C), which can enjoy the cooperative effect of learning both the given classification task and the auxiliary preferences. Here, we provide three different ways to collect preference signals in practice: (a) implicitly extracting from annotation records (for free, but often unavailable), (b) collecting explicitly from crowd workers (high paid), or (c) pre-trained large language models such as GPT-3 (low paid). Given existing classification NLP benchmarks, we demonstrate that the proposed auxiliary preference learning via P2C on them is effective in improving text classifiers. Our codes are publicly available.
****
An Adaptive Entropy-Regularization Framework for Multi-Agent Reinforcement Learning
Woojun Kim, Youngchul Sung
In this paper, we propose an adaptive entropy-regularization framework (ADER) for multi-agent reinforcement learning (RL) to learn the adequate amount of exploration of each agent for entropy-based exploration. In order to derive a metric for the proper level of exploration entropy for each agent, we disentangle the soft value function into two types: one for pure return and the other for entropy. By applying multi-agent value factorization to the disentangled value function of pure return, we obtain a metric to determine the relevant level of exploration entropy for each agent, given by the partial derivative of the pure-return value function with respect to (w.r.t.) the policy entropy of each agent. Based on this metric, we propose the ADER algorithm based on maximum entropy RL, which controls the necessary level of exploration across agents over time by learning the proper target entropy for each agent. Experimental results show that the proposed scheme significantly outperforms current state-of-the-art multi-agent RL algorithms.
****
Practical and Matching Gradient Variance Bounds for Black-Box Variational Bayesian Inference
Kyurae Kim, Kaiwen Wu, Jisu Oh, Jacob R. Gardner
Understanding the gradient variance of black-box variational inference (BBVI) is a crucial step for establishing its convergence and developing algorithmic improvements. However, existing studies have yet to show that the gradient variance of BBVI satisfies the conditions used to study the convergence of stochastic gradient descent (SGD), the workhorse of BBVI. In this work, we show that BBVI satisfies a matching bound corresponding to the ABC condition used in the SGD literature when applied to smooth and quadratically-growing log-likelihoods. Our results generalize to nonlinear covariance parameterizations widely used in the practice of BBVI. Furthermore, we show that the variance of the mean-field parameterization has provably superior dimensional dependence.
****
Learnability and Algorithm for Continual Learning
Gyuhak Kim, Changnan Xiao, Tatsuya Konishi, Bing Liu
This paper studies the challenging continual learning (CL) setting of Class Incr

emental Learning (CIL). CIL learns a sequence of tasks consisting of disjoint se ts of concepts or classes. At any time, a single model is built that can be appl ied to predict/classify test instances of any classes learned thus far without p roviding any task related information for each test instance. Although many tech niques have been proposed for CIL, they are mostly empirical. It has been shown recently that a strong CIL system needs a strong within-task prediction (WP) and a strong out-of-distribution (OOD) detection for each task. However, it is stil l not known whether CIL is actually learnable. This paper shows that CIL is lear nable. Based on the theory, a new CIL algorithm is also proposed. Experimental r esults demonstrate its effectiveness.
****

Unifying Nesterov's Accelerated Gradient Methods for Convex and Strongly Convex Objective Functions
Jungbin Kim, Insoon Yang
Although Nesterov's accelerated gradient method (AGM) has been studied from vari ous perspectives, it remains unclear why the most popular forms of AGMs must han dle convex and strongly convex objective functions separately. To address this i nconsistency, we propose a novel unified framework for Lagrangians, ordinary dif ferential equation (ODE) models, and algorithms. As a special case, our new simp le momentum algorithm, which we call the unified AGM, seamlessly bridges the gap between the two most popular forms of Nesterov's AGM and has a superior converg ence guarantee compared to existing algorithms for non-strongly convex objective functions. This property is beneficial in practice when considering ill-conditi oned μ-strongly convex objective functions (with small μ). Furthermore, we generalize this algorithm and the corresponding ODE model to the higher-order non-Euclidean setting. Last but not least, our unified framework is used to con struct the unified AGM-G ODE, a novel ODE model for minimizing the gradient norm of strongly convex functions.
****

Denoising MCMC for Accelerating Diffusion-Based Generative Models
Beomsu Kim, Jong Chul Ye
The sampling process of diffusion models can be interpreted as solving the rever se stochastic differential equation (SDE) or the ordinary differential equation (ODE) of the diffusion process, which often requires up to thousands of discreti zation steps to generate a single image. This has sparked a great interest in de veloping efficient integration techniques for reverse-S/ODEs. Here, we propose a n orthogonal approach to accelerating score-based sampling: Denoising MCMC (DMCM C). DMCMC first uses MCMC to produce initialization points for reverse-S/ODE in the product space of data and diffusion time. Then, a reverse-S/ODE integrator i s used to denoise the initialization points. Since MCMC traverses close to the d ata manifold, the cost of producing a clean sample for DMCMC is much less than t hat of producing a clean sample from noise. Denoising Langevin Gibbs, an instanc e of DMCMC, successfully accelerates all six reverse-S/ODE integrators considere d in this work, and achieves state-of-the-art results: in the limited number of score function evaluation (NFE) setting on CIFAR10, we have 3.25 FID with $\ap prox 10$ NFE and 2.49 FID with ≈ 16 NFE. On CelebA-HQ-256, we have $6. 99$ FID with ≈ 160 NFE, which beats the current best record of Kim et al . (2022) among score-based models, 7.16 FID with 4000 NFE. Code: https://git hub.com/1202kbs/DMCMC
****

Structure Learning of Latent Factors via Clique Search on Correlation Thresholde d Graphs
Dale Kim, Qing Zhou
Despite the widespread application of latent factor analysis, existing methods s uffer from the following weaknesses: requiring the number of factors to be known , lack of theoretical guarantees for learning the model structure, and nonidenti fiability of the parameters due to rotation invariance properties of the likelih ood. We address these concerns by proposing a fast correlation thresholding (CT) algorithm that simultaneously learns the number of latent factors and a rotatio nally identifiable model structure. Our novel approach translates this structure

learning problem into the search for so-called independent maximal cliques in a thresholded correlation graph that can be easily constructed from the observed data. Our clique analysis technique scales well up to thousands of variables, while competing methods are not applicable in a reasonable amount of running time. We establish a finite-sample error bound and high-dimensional consistency for the structure learning of our method. Through a series of simulation studies and a real data example, we show that the CT algorithm is an accurate method for learning the structure of factor analysis models and is robust to violations of its assumptions.

****

Fair and Robust Estimation of Heterogeneous Treatment Effects for Policy Learning

Kwangho Kim, Jose R Zubizarreta

We propose a simple and general framework for nonparametric estimation of heterogeneous treatment effects under fairness constraints. Under standard regularity conditions, we show that the resulting estimators possess the double robustness property. We use this framework to characterize the trade-off between fairness and the maximum welfare achievable by the optimal policy. We evaluate the methods in a simulation study and illustrate them in a real-world case study.

****

Proper Losses for Discrete Generative Models

Dhamma Kimpara, Rafael Frongillo, Bo Waggoner

We initiate the study of proper losses for evaluating generative models in the discrete setting. Unlike traditional proper losses, we treat both the generative model and the target distribution as black-boxes, only assuming ability to draw i.i.d. samples. We define a loss to be black-box proper if the generative distribution that minimizes expected loss is equal to the target distribution. Using techniques from statistical estimation theory, we give a general construction and characterization of black-box proper losses: they must take a polynomial form, and the number of draws from the model and target distribution must exceed the degree of the polynomial. The characterization rules out a loss whose expectation is the cross-entropy between the target distribution and the model. By extending the construction to arbitrary sampling schemes such as Poisson sampling, however, we show that one can construct such a loss.

****

Controlling Posterior Collapse by an Inverse Lipschitz Constraint on the Decoder Network

Yuri Kinoshita, Kenta Oono, Kenji Fukumizu, Yuichi Yoshida, Shin-Ichi Maeda

Variational autoencoders (VAEs) are one of the deep generative models that have experienced enormous success over the past decades. However, in practice, they suffer from a problem called posterior collapse, which occurs when the posterior distribution coincides, or collapses, with the prior taking no information from the latent structure of the input data into consideration. In this work, we introduce an inverse Lipschitz neural network into the decoder and, based on this architecture, provide a new method that can control in a simple and clear manner the degree of posterior collapse for a wide range of VAE models equipped with a concrete theoretical guarantee. We also illustrate the effectiveness of our method through several numerical experiments.

****

A Watermark for Large Language Models

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, Tom Goldstein

Potential harms of large language models can be mitigated by watermarking model output, i.e., embedding signals into generated text that are invisible to humans but algorithmically detectable from a short span of tokens. We propose a watermarking framework for proprietary language models. The watermark can be embedded with negligible impact on text quality, and can be detected using an efficient open-source algorithm without access to the language model API or parameters. The watermark works by selecting a randomized set of "green" tokens before a word is generated, and then softly promoting use of green tokens during sampling. We p

ropose a statistical test for detecting the watermark with interpretable p-value
s, and derive an information-theoretic framework for analyzing the sensitivity o
f the watermark. We test the watermark using a multi-billion parameter model fro
m the Open Pretrained Transformer (OPT) family, and discuss robustness and secur
ity.
****

Probabilistic Contrastive Learning Recovers the Correct Aleatoric Uncertainty of
 Ambiguous Inputs

Michael Kirchhof, Enkelejda Kasneci, Seong Joon Oh

Contrastively trained encoders have recently been proven to invert the data-gene
rating process: they encode each input, e.g., an image, into the true latent vec
tor that generated the image (Zimmermann et al., 2021). However, real-world obse
rvations often have inherent ambiguities. For instance, images may be blurred or
 only show a 2D view of a 3D object, so multiple latents could have generated th
em. This makes the true posterior for the latent vector probabilistic with heter
oscedastic uncertainty. In this setup, we extend the common InfoNCE objective an
d encoders to predict latent distributions instead of points. We prove that thes
e distributions recover the correct posteriors of the data-generating process, i
ncluding its level of aleatoric uncertainty, up to a rotation of the latent spac
e. In addition to providing calibrated uncertainty estimates, these posteriors a
llow the computation of credible intervals in image retrieval. They comprise ima
ges with the same latent as a given query, subject to its uncertainty. Code is a
t https://github.com/mkirchhof/Probabilistic_Contrastive_Learning .
****

Training Normalizing Flows from Dependent Data

Matthias Kirchler, Christoph Lippert, Marius Kloft

Normalizing flows are powerful non-parametric statistical models that function a
s a hybrid between density estimators and generative models. Current learning al
gorithms for normalizing flows assume that data points are sampled independently
, an assumption that is frequently violated in practice, which may lead to erron
eous density estimation and data generation. We propose a likelihood objective o
f normalizing flows incorporating dependencies between the data points, for whic
h we derive a flexible and efficient learning algorithm suitable for different d
ependency structures. We show that respecting dependencies between observations
can improve empirical results on both synthetic and real-world data, and leads t
o higher statistical power in a downstream application to genome-wide associatio
n studies.
****

IncDSI: Incrementally Updatable Document Retrieval

Varsha Kishore, Chao Wan, Justin Lovelace, Yoav Artzi, Kilian Q Weinberger

Differentiable Search Index is a recently proposed paradigm for document retriev
al, that encodes information about a corpus of documents within the parameters o
f a neural network and directly maps queries to corresponding documents. These m
odels have achieved state-of-the-art performances for document retrieval across
many benchmarks. These kinds of models have a significant limitation: it is not
easy to add new documents after a model is trained. We propose IncDSI, a method
to add documents in real time (about 20-50ms per document), without retraining t
he model on the entire dataset (or even parts thereof). Instead we formulate the
 addition of documents as a constrained optimization problem that makes minimal
changes to the network parameters. Although orders of magnitude faster, our appr
oach is competitive with re-training the model on the whole dataset and enables
the development of document retrieval systems that can be updated with new infor
mation in real-time. Our code for IncDSI is available at https://github.com/vars
hakishore/IncDSI.
****

Regularization and Variance-Weighted Regression Achieves Minimax Optimality in L
inear MDPs: Theory and Practice

Toshinori Kitamura, Tadashi Kozuno, Yunhao Tang, Nino Vieillard, Michal Valko, W
enhao Yang, Jincheng Mei, Pierre Menard, Mohammad Gheshlaghi Azar, Remi Munos, O
livier Pietquin, Matthieu Geist, Csaba Szepesvari, Wataru Kumagai, Yutaka Matsuo

Mirror descent value iteration (MDVI), an abstraction of Kullback-Leibler (KL) and entropy-regularized reinforcement learning (RL), has served as the basis for recent high-performing practical RL algorithms. However, despite the use of function approximation in practice, the theoretical understanding of MDVI has been limited to tabular Markov decision processes (MDPs). We study MDVI with linear function approximation through its sample complexity required to identify an ε-optimal policy with probability $1-\delta$ under the settings of an infinite-horizon linear MDP, generative model, and G-optimal design. We demonstrate that least-squares regression weighted by the variance of an estimated optimal value function of the next state is crucial to achieving minimax optimality. Based on this observation, we present Variance-Weighted Least-Squares MDVI (VWLS-MDVI), the first theoretical algorithm that achieves nearly minimax optimal sample complexity for infinite-horizon linear MDPs. Furthermore, we propose a practical VWLS algorithm for value-based deep RL, Deep Variance Weighting (DVW). Our experiments demonstrate that DVW improves the performance of popular value-based deep RL algorithms on a set of MinAtar benchmarks.
****

Drug Discovery under Covariate Shift with Domain-Informed Prior Distributions over Functions
Leo Klarner, Tim G. J. Rudner, Michael Reutlinger, Torsten Schindler, Garrett M Morris, Charlotte Deane, Yee Whye Teh
Accelerating the discovery of novel and more effective therapeutics is an important pharmaceutical problem in which deep learning is playing an increasingly significant role. However, real-world drug discovery tasks are often characterized by a scarcity of labeled data and significant covariate shift—a setting that poses a challenge to standard deep learning methods. In this paper, we present Q-SAVI, a probabilistic model able to address these challenges by encoding explicit prior knowledge of the data-generating process into a prior distribution over functions, presenting researchers with a transparent and probabilistically principled way to encode data-driven modeling preferences. Building on a novel, gold-standard bioactivity dataset that facilitates a meaningful comparison of models in an extrapolative regime, we explore different approaches to induce data shift and construct a challenging evaluation setup. We then demonstrate that using Q-SAVI to integrate contextualized prior knowledge of drug-like chemical space into the modeling process affords substantial gains in predictive accuracy and calibration, outperforming a broad range of state-of-the-art self-supervised pre-training and domain adaptation techniques.
****

Deep Laplacian-based Options for Temporally-Extended Exploration
Martin Klissarov, Marlos C. Machado
Selecting exploratory actions that generate a rich stream of experience for better learning is a fundamental challenge in reinforcement learning (RL). An approach to tackle this problem consists in selecting actions according to specific policies for an extended period of time, also known as options. A recent line of work to derive such exploratory options builds upon the eigenfunctions of the graph Laplacian. Importantly, until now these methods have been mostly limited to tabular domains where (1) the graph Laplacian matrix was either given or could be fully estimated, (2) performing eigendecomposition on this matrix was computationally tractable, and (3) value functions could be learned exactly. Additionally, these methods required a separate option discovery phase. These assumptions are fundamentally not scalable. In this paper we address these limitations and show how recent results for directly approximating the eigenfunctions of the Laplacian can be leveraged to truly scale up options-based exploration. To do so, we introduce a fully online deep RL algorithm for discovering Laplacian-based options and evaluate our approach on a variety of pixel-based tasks. We compare to several state-of-the-art exploration methods and show that our approach is effective, general, and especially promising in non-stationary settings.
****

Generalized Reductions: Making any Hierarchical Clustering Fair and Balanced with Low Cost

Marina Knittel, Max Springer, John P Dickerson, Mohammadtaghi Hajiaghayi
Clustering is a fundamental building block of modern statistical analysis pipelines. Fair clustering has seen much attention from the machine learning community in recent years. We are some of the first to study fairness in the context of hierarchical clustering, after the results of Ahmadian et al. from NeurIPS in 2020. We evaluate our results using Dasgupta's cost function, perhaps one of the most prevalent theoretical metrics for hierarchical clustering evaluation. Our work vastly improves the previous $O(n^{5/6}poly\log(n))$ fair approximation for cost to a near polylogarithmic $O(n^\delta poly\log(n))$ fair approximation for any constant $\delta\in(0,1)$. This result establishes a cost fairness tradeoff and extends to broader fairness constraints than the previous work. We also show how to alter existing hierarchical clusterings to guarantee fairness and cluster balance across any level in the hierarchy.
****

Can We Scale Transformers to Predict Parameters of Diverse ImageNet Models?
Boris Knyazev, Doha Hwang, Simon Lacoste-Julien
Pretraining a neural network on a large dataset is becoming a cornerstone in machine learning that is within the reach of only a few communities with large-resources. We aim at an ambitious goal of democratizing pretraining. Towards that goal, we train and release a single neural network that can predict high quality ImageNet parameters of other neural networks. By using predicted parameters for initialization we are able to boost training of diverse ImageNet models available in PyTorch. When transferred to other datasets, models initialized with predicted parameters also converge faster and reach competitive final performance.
****

Online Learning with Feedback Graphs: The True Shape of Regret
Tomáš Kocák, Alexandra Carpentier
Sequential learning with feedback graphs is a natural extension of the multi-armed bandit problem where the problem is equipped with an underlying graph structure that provides additional information - playing an action reveals the losses of all the neighbors of the action. This problem was introduced by Mannor & Shamir (2011) and received considerable attention in recent years. It is generally stated in the literature that the minimax regret rate for this problem is of order $\sqrt{\alpha T}$, where α is the independence number of the graph, and T is the time horizon. However, this is proven only when the number of rounds T is larger than α^3, which poses a significant restriction for the usability of this result in large graphs. In this paper, we define a new quantity R^*, called the problem complexity, and prove that the minimax regret is proportional to R^* for any graph and time horizon T. Introducing an intricate exploration strategy, we define the Exp3-EX algorithm that achieves the minimax optimal regret bound and becomes the first provably optimal algorithm for this setting, even if T is smaller than α^3.
****

Grounding Language Models to Images for Multimodal Inputs and Outputs
Jing Yu Koh, Ruslan Salakhutdinov, Daniel Fried
We propose an efficient method to ground pretrained text-only language models to the visual domain, enabling them to process arbitrarily interleaved image-and-text data, and generate text interleaved with retrieved images. Our method leverages the abilities of language models learnt from large scale text-only pretraining, such as in-context learning and free-form text generation. We keep the language model frozen, and finetune input and output linear layers to enable cross-modality interactions. This allows our model to process arbitrarily interleaved image-and-text inputs, and generate free-form text interleaved with retrieved images. We achieve strong zero-shot performance on grounded tasks such as contextual image retrieval and multimodal dialogue, and showcase compelling interactive abilities. Our approach works with any off-the-shelf language model and paves the way towards an effective, general solution for leveraging pretrained language models in visually grounded settings.
****

Rigid Body Flows for Sampling Molecular Crystal Structures

Jonas Köhler, Michele Invernizzi, Pim De Haan, Frank Noe
Normalizing flows (NF) are a class of powerful generative models that have gained popularity in recent years due to their ability to model complex distributions with high flexibility and expressiveness. In this work, we introduce a new type of normalizing flow that is tailored for modeling positions and orientations of multiple objects in three-dimensional space, such as molecules in a crystal. Our approach is based on two key ideas: first, we define smooth and expressive flows on the group of unit quaternions, which allows us to capture the continuous rotational motion of rigid bodies; second, we use the double cover property of unit quaternions to define a proper density on the rotation group. This ensures that our model can be trained using standard likelihood-based methods or variational inference with respect to a thermodynamic target density. We evaluate the method by training Boltzmann generators for two molecular examples, namely the multi-modal density of a tetrahedral system in an external field and the ice XI phase in the TIP4P water model. Our flows can be combined with flows operating on the internal degrees of freedom of molecules and constitute an important step towards the modeling of distributions of many interacting molecules.
****

Enabling First-Order Gradient-Based Learning for Equilibrium Computation in Markets
Nils Kohring, Fabian Raoul Pieroth, Martin Bichler
Understanding and analyzing markets is crucial, yet analytical equilibrium solutions remain largely infeasible. Recent breakthroughs in equilibrium computation rely on zeroth-order policy gradient estimation. These approaches commonly suffer from high variance and are computationally expensive. The use of fully differentiable simulators would enable more efficient gradient estimation. However, the discrete allocation of goods in economic simulations is a non-differentiable operation. This renders the first-order Monte Carlo gradient estimator inapplicable and the learning feedback systematically misleading. We propose a novel smoothing technique that creates a surrogate market game, in which first-order methods can be applied. We provide theoretical bounds on the resulting bias which justifies solving the smoothed game instead. These bounds also allow choosing the smoothing strength a priori such that the resulting estimate has low variance. Furthermore, we validate our approach via numerous empirical experiments. Our method theoretically and empirically outperforms zeroth-order methods in approximation quality and computational efficiency.
****

Revisiting Gradient Clipping: Stochastic bias and tight convergence guarantees
Anastasia Koloskova, Hadrien Hendrikx, Sebastian U Stich
Gradient clipping is a popular modification to standard (stochastic) gradient descent, at every iteration limiting the gradient norm to a certain value $c >0$. It is widely used for example for stabilizing the training of deep learning models (Goodfellow et al., 2016), or for enforcing differential privacy (Abadi et al., 2016). Despite popularity and simplicity of the clipping mechanism, its convergence guarantees often require specific values of c and strong noise assumptions. In this paper, we give convergence guarantees that show precise dependence on arbitrary clipping thresholds c and show that our guarantees are tight with both deterministic and stochastic gradients. In particular, we show that (i) for deterministic gradient descent, the clipping threshold only affects the higher-order terms of convergence, (ii) in the stochastic setting convergence to the true optimum cannot be guaranteed under the standard noise assumption, even under arbitrary small step-sizes. We give matching upper and lower bounds for convergence of the gradient norm when running clipped SGD, and illustrate these results with experiments.
****

On Computing Optimal Tree Ensembles
Christian Komusiewicz, Pascal Kunz, Frank Sommer, Manuel Sorge
Random forests and, more generally, (decision-)tree ensembles are widely used methods for classification and regression. Recent algorithmic advances allow to compute decision trees that are optimal for various measures such as their size or

depth. We are not aware of such research for tree ensembles and aim to contribu
te to this area. Mainly, we provide two novel algorithms and corresponding lower
bounds. First, we are able to carry over and substantially improve on tractabil
ity results for decision trees, obtaining a $(6\delta D S)^S \cdot \mathrm{poly}
$-time algorithm, where S is the number of cuts in the tree ensemble, D the
largest domain size, and δ is the largest number of features in which two
examples differ. To achieve this, we introduce the witness-tree technique which
also seems promising for practice. Second, we show that dynamic programming, wh
ich has been successful for decision trees, may also be viable for tree ensemble
s, providing an $\ell^n \cdot \mathrm{poly}$-time algorithm, where ℓ is the
number of trees and n the number of examples. Finally, we compare the number
of cuts necessary to classify training data sets for decision trees and tree ens
embles, showing that ensembles may need exponentially fewer cuts for increasing
number of trees.
****

GOAT: A Global Transformer on Large-scale Graphs
Kezhi Kong, Jiuhai Chen, John Kirchenbauer, Renkun Ni, C. Bayan Bruss, Tom Golds
tein
Graph transformers have been competitive on graph classification tasks, but they
fail to outperform Graph Neural Networks (GNNs) on node classification, which i
s a common task performed on large-scale graphs for industrial applications. Mea
nwhile, existing GNN architectures are limited in their ability to perform equal
ly well on both homophilious and heterophilious graphs as their inductive biases
are generally tailored to only one setting. To address these issues, we propose
GOAT, a scalable global graph transformer. In GOAT, each node conceptually atte
nds to all the nodes in the graph and homophily/heterophily relationships can be
learnt adaptively from the data. We provide theoretical justification for our a
pproximate global self-attention scheme, and show it to be scalable to large-sca
le graphs. We demonstrate the competitiveness of GOAT on both heterophilious and
homophilious graphs with millions of nodes.
****

Autoregressive Diffusion Model for Graph Generation
Lingkai Kong, Jiaming Cui, Haotian Sun, Yuchen Zhuang, B. Aditya Prakash, Chao Z
hang
Diffusion-based graph generative models have recently obtained promising results
for graph generation. However, existing diffusion-based graph generative models
are mostly one-shot generative models that apply Gaussian diffusion in the dequ
antized adjacency matrix space. Such a strategy can suffer from difficulty in mo
del training, slow sampling speed, and incapability of incorporating constraints
. We propose an autoregressive diffusion model for graph generation. Unlike exis
ting methods, we define a node-absorbing diffusion process that operates directl
y in the discrete graph space. For forward diffusion, we design a diffusion orde
ring network, which learns a data-dependent node absorbing ordering from graph t
opology. For reverse generation, we design a denoising network that uses the rev
erse node ordering to efficiently reconstruct the graph by predicting the node t
ype of the new node and its edges with previously denoised nodes at a time. Base
d on the permutation invariance of graph, we show that the two networks can be j
ointly trained by optimizing a simple lower bound of data likelihood. Our experi
ments on six diverse generic graph datasets and two molecule datasets show that
our model achieves better or comparable generation performance with previous sta
te-of-the-art, and meanwhile enjoys fast generation speed.
****

End-to-End Full-Atom Antibody Design
Xiangzhe Kong, Wenbing Huang, Yang Liu
Antibody design is an essential yet challenging task in various domains like the
rapeutics and biology. There are two major defects in current learning-based met
hods: 1) tackling only a certain subtask of the whole antibody design pipeline,
making them suboptimal or resource-intensive. 2) omitting either the framework r
egions or side chains, thus incapable of capturing the full-atom geometry. To ad
dress these pitfalls, we propose dynamic Multi-channel Equivariant grAph Network

(dyMEAN), an end-to-end full-atom model for E(3)-equivariant antibody design gi
ven the epitope and the incomplete sequence of the antibody. Specifically, we fi
rst explore structural initialization as a knowledgeable guess of the antibody s
tructure and then propose shadow paratope to bridge the epitope-antibody connect
ions. Both 1D sequences and 3D structures are updated via an adaptive multi-chan
nel equivariant encoder that is able to process protein residues of variable siz
es when considering full atoms. Finally, the updated antibody is docked to the e
pitope via the alignment of the shadow paratope. Experiments on epitope-binding
CDR-H3 design, complex structure prediction, and affinity optimization demonstra
te the superiority of our end-to-end framework and full-atom modeling.
****

Covariate balancing using the integral probability metric for causal inference
Insung Kong, Yuha Park, Joonhyuk Jung, Kwonsang Lee, Yongdai Kim
Weighting methods in causal inference have been widely used to achieve a desirab
le level of covariate balancing. However, the existing weighting methods have de
sirable theoretical properties only when a certain model, either the propensity
score or outcome regression model, is correctly specified. In addition, the corr
esponding estimators do not behave well for finite samples due to large variance
even when the model is correctly specified. In this paper, we consider to use t
he integral probability metric (IPM), which is a metric between two probability
measures, for covariate balancing. Optimal weights are determined so that weight
ed empirical distributions for the treated and control groups have the smallest
IPM value for a given set of discriminators. We prove that the corresponding est
imator can be consistent without correctly specifying any model (neither the pro
pensity score nor the outcome regression model). In addition, we empirically sho
w that our proposed method outperforms existing weighting methods with large mar
gins for finite samples.
****

Masked Bayesian Neural Networks : Theoretical Guarantee and its Posterior Infere
nce
Insung Kong, Dongyoon Yang, Jongjin Lee, Ilsang Ohn, Gyuseung Baek, Yongdai Kim
Bayesian approaches for learning deep neural networks (BNN) have been received m
uch attention and successfully applied to various applications. Particularly, BN
Ns have the merit of having better generalization ability as well as better unce
rtainty quantification. For the success of BNN, search an appropriate architectu
re of the neural networks is an important task, and various algorithms to find g
ood sparse neural networks have been proposed. In this paper, we propose a new n
ode-sparse BNN model which has good theoretical properties and is computationall
y feasible. We prove that the posterior concentration rate to the true model is
near minimax optimal and adaptive to the smoothness of the true model. In partic
ular the adaptiveness is the first of its kind for node-sparse BNNs. In addition
, we develop a novel MCMC algorithm which makes the Bayesian inference of the no
de-sparse BNN model feasible in practice.
****

Parameter-Level Soft-Masking for Continual Learning
Tatsuya Konishi, Mori Kurokawa, Chihiro Ono, Zixuan Ke, Gyuhak Kim, Bing Liu
Existing research on task incremental learning in continual learning has primari
ly focused on preventing catastrophic forgetting (CF). Although several techniqu
es have achieved learning with no CF, they attain it by letting each task monopo
lize a sub-network in a shared network, which seriously limits knowledge transfe
r (KT) and causes over-consumption of the network capacity, i.e., as more tasks
are learned, the performance deteriorates. The goal of this paper is threefold:
(1) overcoming CF, (2) encouraging KT, and (3) tackling the capacity problem. A
novel technique (called SPG) is proposed that soft-masks (partially blocks) para
meter updating in training based on the importance of each parameter to old task
s. Each task still uses the full network, i.e., no monopoly of any part of the n
etwork by any task, which enables maximum KT and reduction in capacity usage. To
our knowledge, this is the first work that soft-masks a model at the parameter-
level for continual learning. Extensive experiments demonstrate the effectivenes
s of SPG in achieving all three objectives. More notably, it attains significant

transfer of knowledge not only among similar tasks (with shared knowledge) but also among dissimilar tasks (with little shared knowledge) while mitigating CF.
****
Pretraining Language Models with Human Preferences
Tomasz Korbak, Kejian Shi, Angelica Chen, Rasika Vinayak Bhalerao, Christopher Buckley, Jason Phang, Samuel R. Bowman, Ethan Perez
Language models (LMs) are pretrained to imitate text from large and diverse datasets that contain content that would violate human preferences if generated by an LM: falsehoods, offensive comments, personally identifiable information, low-quality or buggy code, among others. Here, we explore alternative objectives for pretraining LMs in a way that also guides them to generate text aligned with human preferences. We benchmark five objectives for pretraining with human feedback across three tasks and study how they affect the alignment and capabilities of pretrained LMs. We find a Pareto-optimal and simple approach among those we explored: conditional training, or learning distribution over tokens conditional on their human preference scores. Conditional training reduces the rate of undesirable content by up to an order of magnitude, both when generating without a prompt and with an adversarially-chosen prompt. Moreover, conditional training maintains the downstream task performance of standard LM pretraining, both before and after task-specific finetuning. Pretraining with human feedback results in much better preference satisfaction than standard LM pretraining followed by finetuning with feedback, i.e., learning and then unlearning undesirable behavior. Our results suggest that we should move beyond imitation learning when pretraining LMs and incorporate human preferences from the start of training.
****
Detecting Adversarial Directions in Deep Reinforcement Learning to Make Robust Decisions
Ezgi Korkmaz, Jonah Brown-Cohen
Learning in MDPs with highly complex state representations is currently possible due to multiple advancements in reinforcement learning algorithm design. However, this incline in complexity, and furthermore the increase in the dimensions of the observation came at the cost of volatility that can be taken advantage of via adversarial attacks (i.e. moving along worst-case directions in the observation space). To solve this policy instability problem we propose a novel method to detect the presence of these non-robust directions via local quadratic approximation of the deep neural policy loss. Our method provides a theoretical basis for the fundamental cut-off between safe observations and adversarial observations. Furthermore, our technique is computationally efficient, and does not depend on the methods used to produce the worst-case directions. We conduct extensive experiments in the Arcade Learning Environment with several different adversarial attack techniques. Most significantly, we demonstrate the effectiveness of our approach even in the setting where non-robust directions are explicitly optimized to circumvent our proposed method.
****
Ewald-based Long-Range Message Passing for Molecular Graphs
Arthur Kosmala, Johannes Gasteiger, Nicholas Gao, Stephan Günnemann
Neural architectures that learn potential energy surfaces from molecular data have undergone fast improvement in recent years. A key driver of this success is the Message Passing Neural Network (MPNN) paradigm. Its favorable scaling with system size partly relies upon a spatial distance limit on messages. While this focus on locality is a useful inductive bias, it also impedes the learning of long-range interactions such as electrostatics and van der Waals forces. To address this drawback, we propose Ewald message passing: a nonlocal Fourier space scheme which limits interactions via a cutoff on frequency instead of distance, and is theoretically well-founded in the Ewald summation method. It can serve as an augmentation on top of existing MPNN architectures as it is computationally inexpensive and agnostic to architectural details. We test the approach with four baseline models and two datasets containing diverse periodic (OC20) and aperiodic structures (OE62). Across all models and datasets, we observe robust improvements in energy mean absolute errors, averaging 10% on OC20 and 16% on OE62. Our analy

sis shows an outsize impact of these improvements on structures with high long-r
ange contributions to the ground-truth energy.
****

TabDDPM: Modelling Tabular Data with Diffusion Models
Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, Artem Babenko
Denoising diffusion probabilistic models are becoming the leading generative mod
eling paradigm for many important data modalities. Being the most prevalent in t
he computer vision community, diffusion models have recently gained some attenti
on in other domains, including speech, NLP, and graph-like data. In this work, w
e investigate if the framework of diffusion models can be advantageous for gener
al tabular problems, where data points are typically represented by vectors of h
eterogeneous features. The inherent heterogeneity of tabular data makes it quite
 challenging for accurate modeling since the individual features can be of a com
pletely different nature, i.e., some of them can be continuous and some can be d
iscrete. To address such data types, we introduce TabDDPM — a diffusion model th
at can be universally applied to any tabular dataset and handles any feature typ
es. We extensively evaluate TabDDPM on a wide set of benchmarks and demonstrate
its superiority over existing GAN/VAE alternatives, which is consistent with the
 advantage of diffusion models in other fields.
****

Randomized Schur Complement Views for Graph Contrastive Learning
Vignesh Kothapalli
We introduce a randomized topological augmentor based on Schur complements for G
raph Contrastive Learning (GCL). Given a graph laplacian matrix, the technique g
enerates unbiased approximations of its Schur complements and treats the corresp
onding graphs as augmented views. We discuss the benefits of our approach, provi
de theoretical justifications and present connections with graph diffusion. Unli
ke previous efforts, we study the empirical effectiveness of the augmentor in a
controlled fashion by varying the design choices for subsequent GCL phases, such
 as encoding and contrasting. Extensive experiments on node and graph classifica
tion benchmarks demonstrate that our technique consistently outperforms pre-defi
ned and adaptive augmentation approaches to achieve state-of-the-art results.
****

Benign Overfitting in Two-layer ReLU Convolutional Neural Networks
Yiwen Kou, Zixiang Chen, Yuanzhou Chen, Quanquan Gu
Modern deep learning models with great expressive power can be trained to overfi
t the training data but still generalize well. This phenomenon is referred to as
 benign overfitting. Recently, a few studies have attempted to theoretically und
erstand benign overfitting in neural networks. However, these works are either l
imited to neural networks with smooth activation functions or to the neural tang
ent kernel regime. How and when benign overfitting can occur in ReLU neural netw
orks remains an open problem. In this work, we seek to answer this question by e
stablishing algorithm-dependent risk bounds for learning two-layer ReLU convolut
ional neural networks with label-flipping noise. We show that, under mild condit
ions, the neural network trained by gradient descent can achieve near-zero train
ing loss and Bayes optimal test risk. Our result also reveals a sharp transition
 between benign and harmful overfitting under different conditions on data distr
ibution in terms of test risk. Experiments on synthetic data back up our theory.
****

Variational Mixture of HyperGenerators for Learning Distributions over Functions
Batuhan Koyuncu, Pablo Sanchez Martin, Ignacio Peis, Pablo M. Olmos, Isabel Vale
ra
Recent approaches build on implicit neural representations (INRs) to propose gen
erative models over function spaces. However, they are computationally costly wh
en dealing with inference tasks, such as missing data imputation, or directly ca
nnot tackle them. In this work, we propose a novel deep generative model, named
VaMoH. VaMoH combines the capabilities of modeling continuous functions using IN
Rs and the inference capabilities of Variational Autoencoders (VAEs). In additio
n, VaMoH relies on a normalizing flow to define the prior, and a mixture of hype
rnetworks to parametrize the data log-likelihood. This gives VaMoH a high expres

sive capability and interpretability. Through experiments on a diverse range of data types, such as images, voxels, and climate data, we show that VaMoH can effectively learn rich distributions over continuous functions. Furthermore, it can perform inference-related tasks, such as conditional super-resolution generation and in-painting, as well or better than previous approaches, while being less computationally demanding.
****

Gradient Descent Monotonically Decreases the Sharpness of Gradient Flow Solutions in Scalar Networks and Beyond
Itai Kreisler, Mor Shpigel Nacson, Daniel Soudry, Yair Carmon
Recent research shows that when Gradient Descent (GD) is applied to neural networks, the loss almost never decreases monotonically. Instead, the loss oscillates as gradient descent converges to its “Edge of Stability” (EoS). Here, we find a quantity that does decrease monotonically throughout GD training: the sharpness attained by the gradient flow solution (GFS)—the solution that would be obtained if, from now until convergence, we train with an infinitesimal step size. Theoretically, we analyze scalar neural networks with the squared loss, perhaps the simplest setting where the EoS phenomena still occur. In this model, we prove that the GFS sharpness decreases monotonically. Using this result, we characterize settings where GD provably converges to the EoS in scalar networks. Empirically, we show that GD monotonically decreases the GFS sharpness in a squared regression model as well as practical neural network architectures.
****

Estimation Beyond Data Reweighting: Kernel Method of Moments
Heiner Kremer, Yassine Nemmour, Bernhard Schölkopf, Jia-Jie Zhu
Moment restrictions and their conditional counterparts emerge in many areas of machine learning and statistics ranging from causal inference to reinforcement learning. Estimators for these tasks, generally called methods of moments, include the prominent generalized method of moments (GMM) which has recently gained attention in causal inference. GMM is a special case of the broader family of empirical likelihood estimators which are based on approximating a population distribution by means of minimizing a φ-divergence to an empirical distribution. However, the use of φ-divergences effectively limits the candidate distributions to reweightings of the data samples. We lift this long-standing limitation and provide a method of moments that goes beyond data reweighting. This is achieved by defining an empirical likelihood estimator based on maximum mean discrepancy which we term the kernel method of moments (KMM). We provide a variant of our estimator for conditional moment restrictions and show that it is asymptotically first-order optimal for such problems. Finally, we show that our method achieves competitive performance on several conditional moment restriction tasks.
****

Multi-Task Differential Privacy Under Distribution Skew
Walid Krichene, Prateek Jain, Shuang Song, Mukund Sundararajan, Abhradeep Guha Thakurta, Li Zhang
We study the problem of multi-task learning under user-level differential privacy, in which n users contribute data to m tasks, each involving a subset of users. One important aspect of the problem, that can significantly impact quality, is the distribution skew among tasks. Tasks that have much fewer data samples than others are more susceptible to the noise added for privacy. It is natural to ask whether algorithms can adapt to this skew to improve the overall utility. We give a systematic analysis of the problem, by studying how to optimally allocate a user’s privacy budget among tasks. We propose a generic algorithm, based on an adaptive reweighting of the empirical loss, and show that in the presence of distribution skew, this gives a quantifiable improvement of excess empirical risk. Experimental studies on recommendation problems that exhibit a long tail of small tasks, demonstrate that our methods significantly improve utility, achieving the state of the art on two standard benchmarks.
****

Towards Bridging the Gaps between the Right to Explanation and the Right to be F

orgotten
Satyapriya Krishna, Jiaqi Ma, Himabindu Lakkaraju
The Right to Explanation and the Right to be Forgotten are two important principles outlined to regulate algorithmic decision making and data usage in real-world applications. While the right to explanation allows individuals to request an actionable explanation for an algorithmic decision, the right to be forgotten grants them the right to ask for their data to be deleted from all the databases and models of an organization. Intuitively, enforcing the right to be forgotten may trigger model updates which in turn invalidate previously provided explanations, thus violating the right to explanation. In this work, we investigate the technical implications arising due to the interference between the two aforementioned regulatory principles, and propose the first algorithmic framework to resolve the tension between them. To this end, we formulate a novel optimization problem to generate explanations that are robust to model updates due to the removal of training data instances by data deletion requests. We then derive an efficient approximation algorithm to handle the combinatorial complexity of this optimization problem. We theoretically demonstrate that our method generates explanations that are provably robust to worst-case data deletion requests with bounded costs in case of linear models and certain classes of non-linear models. Extensive experimentation with real-world datasets demonstrates the efficacy of the proposed framework.
****
Graph Neural Tangent Kernel: Convergence on Large Graphs
Sanjukta Krishnagopal, Luana Ruiz
Graph neural networks (GNNs) achieve remarkable performance in graph machine learning tasks but can be hard to train on large-graph data, where their learning dynamics are not well understood. We investigate the training dynamics of large-graph GNNs using graph neural tangent kernels (GNTKs) and graphons. In the limit of large width, optimization of an overparametrized NN is equivalent to kernel regression on the NTK. Here, we investigate how the GNTK evolves as another independent dimension is varied: the graph size. We use graphons to define limit objects—graphon NNs for GNNs, and graphon NTKs for GNTKs—, and prove that, on a sequence of graphs, the GNTKs converge to the graphon NTK. We further prove that the spectrum of the GNTK, which is related to the problem’s learning directions, converges to the spectrum of the GNTK. This implies that in the large-graph limit, the GNTK fitted on a graph of moderate size can be used to solve the same task on the large graph, and to infer the learning dynamics of the large-graph GNN. These results are verified empirically on node regression and classification tasks.
****
Diffusion Models for Black-Box Optimization
Siddarth Krishnamoorthy, Satvik Mehul Mashkaria, Aditya Grover
The goal of offline black-box optimization (BBO) is to optimize an expensive black-box function using a fixed dataset of function evaluations. Prior works consider forward approaches that learn surrogates to the black-box function and inverse approaches that directly map function values to corresponding points in the input domain of the black-box function. These approaches are limited by the quality of the offline dataset and the difficulty in learning one-to-many mappings in high dimensions, respectively. We propose Denoising Diffusion Optimization Models (DDOM), a new inverse approach for offline black-box optimization based on diffusion models. Given an offline dataset, DDOM learns a conditional generative model over the domain of the black-box function conditioned on the function values. We investigate several design choices in DDOM, such as reweighting the dataset to focus on high function values and the use of classifier-free guidance at test-time to enable generalization to function values that can even exceed the dataset maxima. Empirically, we conduct experiments on the Design-Bench benchmark (Trabucco et al., 2022) and show that DDOM achieves results competitive with state-of-the-art baselines.
****
Learning to Design Analog Circuits to Meet Threshold Specifications

Dmitrii Krylov, Pooya Khajeh, Junhan Ouyang, Thomas Reeves, Tongkai Liu, Hiba Ajmal, Hamidreza Aghasi, Roy Fox
Automated design of analog and radio-frequency circuits using supervised or reinforcement learning from simulation data has recently been studied as an alternative to manual expert design. It is straightforward for a design agent to learn an inverse function from desired performance metrics to circuit parameters. However, it is more common for a user to have threshold performance criteria rather than an exact target vector of feasible performance measures. In this work, we propose a method for generating from simulation data a dataset on which a system can be trained via supervised learning to design circuits to meet threshold specifications. We moreover perform the to-date most extensive evaluation of automated analog circuit design, including experimenting in a significantly more diverse set of circuits than in prior work, covering linear, nonlinear, and autonomous circuit configurations, and show that our method consistently reaches success rate better than 90% at 5% error margin, while also improving data efficiency by upward of an order of magnitude.
****

Variance Control for Distributional Reinforcement Learning
Qi Kuang, Zhoufan Zhu, Liwen Zhang, Fan Zhou
Although distributional reinforcement learning (DRL) has been widely examined in the past few years, very few studies investigate the validity of the obtained Q-function estimator in the distributional setting. To fully understand how the approximation errors of the Q-function affect the whole training process, we do some error analysis and theoretically show how to reduce both the bias and the variance of the error terms. With this new understanding, we construct a new estimator Quantiled Expansion Mean (QEM) and introduce a new DRL algorithm (QEMRL) from the statistical perspective. We extensively evaluate our QEMRL algorithm on a variety of Atari and Mujoco benchmark tasks and demonstrate that QEMRL achieves significant improvement over baseline algorithms in terms of sample efficiency and convergence performance.
****

Hierarchical Imitation Learning with Vector Quantized Models
Kalle Kujanpää, Joni Pajarinen, Alexander Ilin
The ability to plan actions on multiple levels of abstraction enables intelligent agents to solve complex tasks effectively. However, learning the models for both low and high-level planning from demonstrations has proven challenging, especially with higher-dimensional inputs. To address this issue, we propose to use reinforcement learning to identify subgoals in expert trajectories by associating the magnitude of the rewards with the predictability of low-level actions given the state and the chosen subgoal. We build a vector-quantized generative model for the identified subgoals to perform subgoal-level planning. In experiments, the algorithm excels at solving complex, long-horizon decision-making problems outperforming state-of-the-art. Because of its ability to plan, our algorithm can find better trajectories than the ones in the training set.
****

SinDDM: A Single Image Denoising Diffusion Model
Vladimir Kulikov, Shahar Yadin, Matan Kleiner, Tomer Michaeli
Denoising diffusion models (DDMs) have led to staggering performance leaps in image generation, editing and restoration. However, existing DDMs use very large datasets for training. Here, we introduce a framework for training a DDM on a single image. Our method, which we coin SinDDM, learns the internal statistics of the training image by using a multi-scale diffusion process. To drive the reverse diffusion process, we use a fully-convolutional light-weight denoiser, which is conditioned on both the noise level and the scale. This architecture allows generating samples of arbitrary dimensions, in a coarse-to-fine manner. As we illustrate, SinDDM generates diverse high-quality samples, and is applicable in a wide array of tasks, including style transfer and harmonization. Furthermore, it can be easily guided by external supervision. Particularly, we demonstrate text-guided generation from a single image using a pre-trained CLIP model.
****

Towards Explaining Distribution Shifts
Sean Kulinski, David I. Inouye
A distribution shift can have fundamental consequences such as signaling a change in the operating environment or significantly reducing the accuracy of downstream models. Thus, understanding distribution shifts is critical for examining and hopefully mitigating the effect of such a shift. Most prior work has focused on merely detecting if a shift has occurred and assumes any detected shift can be understood and handled appropriately by a human operator. We hope to aid in these manual mitigation tasks by explaining the distribution shift using interpretable transportation maps from the original distribution to the shifted one. We derive our interpretable mappings from a relaxation of the optimal transport problem, where the candidate mappings are restricted to a set of interpretable mappings. We then use a wide array of quintessential examples of distribution shift in real-world tabular, text, and image cases to showcase how our explanatory mappings provide a better balance between detail and interpretability than baseline explanations by both visual inspection and our PercentExplained metric.
****

Featured Graph Coarsening with Similarity Guarantees
Manoj Kumar, Anurag Sharma, Shashwat Saxena, Sandeep Kumar
Graph coarsening is a dimensionality reduction technique that aims to learn a smaller-tractable graph while preserving the properties of the original input graph. However, many real-world graphs also have features or contexts associated with each node. The existing graph coarsening methods do not consider the node features and rely solely on a graph matrix(e.g., adjacency and Laplacian) to coarsen graphs. However, some recent deep learning-based graph coarsening methods are designed for specific tasks considering both node features and graph matrix. In this paper, we introduce a novel optimization-based framework for graph coarsening that takes both the graph matrix and the node features as the input and jointly learns the coarsened graph matrix and the coarsened feature matrix while ensuring desired properties. To the best of our knowledge, this is the first work that guarantees that the learned coarsened graph is $\epsilon\in[0,1)$ similar to the original graph. Extensive experiments with both real and synthetic benchmark datasets elucidate the proposed framework’s efficacy and applicability for numerous graph-based applications, including graph clustering, node classification, stochastic block model identification, and graph summarization.
****

Modeling Dynamic Environments with Scene Graph Memory
Andrey Kurenkov, Michael Lingelbach, Tanmay Agarwal, Emily Jin, Chengshu Li, Ruohan Zhang, Li Fei-Fei, Jiajun Wu, Silvio Savarese, Roberto Mart■■n-Mart■■n
Embodied AI agents that search for objects in large environments such as households often need to make efficient decisions by predicting object locations based on partial information. We pose this as a new type of link prediction problem: link prediction on partially observable dynamic graphs Our graph is a representation of a scene in which rooms and objects are nodes, and their relationships are encoded in the edges; only parts of the changing graph are known to the agent at each timestep. This partial observability poses a challenge to existing link prediction approaches, which we address. We propose a novel state representation – Scene Graph Memory (SGM) – with captures the agent’s accumulated set of observations, as well as a neural net architecture called a Node Edge Predictor (NEP) that extracts information from the SGM to search efficiently. We evaluate our method in the Dynamic House Simulator, a new benchmark that creates diverse dynamic graphs following the semantic patterns typically seen at homes, and show that NEP can be trained to predict the locations of objects in a variety of environments with diverse object movement dynamics, outperforming baselines both in terms of new scene adaptability and overall accuracy. The codebase and more can be found www.scenegraphmemory.com.
****

Tied-Augment: Controlling Representation Similarity Improves Data Augmentation
Emirhan Kurtulu■, Zichao Li, Yann Dauphin, Ekin Dogus Cubuk
Data augmentation methods have played an important role in the recent advance of

deep learning models, and have become an indispensable component of state-of-the-art models in semi-supervised, self-supervised, and supervised training for vision. Despite incurring no additional latency at test time, data augmentation often requires more epochs of training to be effective. For example, even the simple flips-and-crops augmentation requires training for more than 5 epochs to improve performance, whereas RandAugment requires more than 90 epochs. We propose a general framework called Tied-Augment, which improves the efficacy of data augmentation in a wide range of applications by adding a simple term to the loss that can control the similarity of representations under distortions. Tied-Augment can improve state-of-the-art methods from data augmentation (e.g. RandAugment, mixup), optimization (e.g. SAM), and semi-supervised learning (e.g. FixMatch). For example, Tied-RandAugment can outperform RandAugment by 2.0% on ImageNet. Notably, using Tied-Augment, data augmentation can be made to improve generalization even when training for a few epochs and when fine-tuning. We open source our code at https://github.com/ekurtulus/tied-augment/tree/main.

****

Cooperation in the Latent Space: The Benefits of Adding Mixture Components in Variational Autoencoders

Oskar Kviman, Ricky Molén, Alexandra Hotti, Semih Kurt, V■■ctor Elvira, Jens Lagergren

In this paper, we show how the mixture components cooperate when they jointly adapt to maximize the ELBO. We build upon recent advances in the multiple and adaptive importance sampling literature. We then model the mixture components using separate encoder networks and show empirically that the ELBO is monotonically non-decreasing as a function of the number of mixture components. These results hold for a range of different VAE architectures on the MNIST, FashionMNIST, and CIFAR-10 datasets. In this work, we also demonstrate that increasing the number of mixture components improves the latent-representation capabilities of the VAE on both image and single-cell datasets. This cooperative behavior motivates that using Mixture VAEs should be considered a standard approach for obtaining more flexible variational approximations. Finally, Mixture VAEs are here, for the first time, compared and combined with normalizing flows, hierarchical models and/or the VampPrior in an extensive ablation study. Multiple of our Mixture VAEs achieve state-of-the-art log-likelihood results for VAE architectures on the MNIST and FashionMNIST datasets. The experiments are reproducible using our code, provided https://github.com/Lagergren-Lab/MixtureVAEs.

****

GeCoNeRF: Few-shot Neural Radiance Fields via Geometric Consistency

Min-Seop Kwak, Jiuhn Song, Seungryong Kim

We present a novel framework to regularize Neural Radiance Field (NeRF) in a few-shot setting with a geometry-aware consistency regularization. The proposed approach leverages a rendered depth map at unobserved viewpoint to warp sparse input images to the unobserved viewpoint and impose them as pseudo ground truths to facilitate learning of NeRF. By encouraging such geometry-aware consistency at a feature-level instead of using pixel-level reconstruction loss, we regularize the NeRF at semantic and structural levels while allowing for modeling view dependent radiance to account for color variations across viewpoints. We also propose an effective method to filter out erroneous warped solutions, along with training strategies to stabilize training during optimization. We show that our model achieves competitive results compared to state-of-the-art few-shot NeRF models.

****

Rotation and Translation Invariant Representation Learning with Implicit Neural Representations

Sehyun Kwon, Joo Young Choi, Ernest K. Ryu

In many computer vision applications, images are acquired with arbitrary or random rotations and translations, and in such setups, it is desirable to obtain semantic representations disentangled from the image orientation. Examples of such applications include semiconductor wafer defect inspection, plankton microscope images, and inference on single-particle cryo-electron microscopy (cryo-EM) micrographs. In this work, we propose Invariant Representation Learning with Implic

it Neural Representation (IRL-INR), which uses an implicit neural representation (INR) with a hypernetwork to obtain semantic representations disentangled from the orientation of the image. We show that IRL-INR can effectively learn disentangled semantic representations on more complex images compared to those considered in prior works and show that these semantic representations synergize well with SCAN to produce state-of-the-art unsupervised clustering results.
****

Reward-Mixing MDPs with Few Latent Contexts are Learnable
Jeongyeol Kwon, Yonathan Efroni, Constantine Caramanis, Shie Mannor
We consider episodic reinforcement learning in reward-mixing Markov decision processes (RMMDPs): at the beginning of every episode nature randomly picks a latent reward model among M candidates and an agent interacts with the MDP throughout the episode for H time steps. Our goal is to learn a near-optimal policy that nearly maximizes the H time-step cumulative rewards in such a model. Prior work established an upper bound for RMMDPs with $M=2$. In this work, we resolve several open questions for the general RMMDP setting. We consider an arbitrary $M\ge2$ and provide a sample-efficient algorithm–EM^2–that outputs an ϵ-optimal policy using $O \left(\epsilon^{-2} \cdot S^d A^d \cdot \text{poly}(H, Z)^d \right)$ episodes, where S, A are the number of states and actions respectively, H is the time-horizon, Z is the support size of reward distributions and $d=O(\min(M,H))$. We also provide a $(SA)^{\Omega(\sqrt{M})} / \epsilon^{2}$ lower bound, supporting that super-polynomial sample complexity in M is necessary.
****

A Fully First-Order Method for Stochastic Bilevel Optimization
Jeongyeol Kwon, Dohyun Kwon, Stephen Wright, Robert D Nowak
We consider stochastic unconstrained bilevel optimization problems when only the first-order gradient oracles are available. While numerous optimization methods have been proposed for tackling bilevel problems, existing methods either tend to require possibly expensive calculations regarding Hessians of lower-level objectives, or lack rigorous finite-time performance guarantees. In this work, we propose a Fully First-order Stochastic Approximation (F2SA) method, and study its non-asymptotic convergence properties. Specifically, we show that F2SA converges to an ϵ-stationary solution of the bilevel problem after $\epsilon^{-7/2}, \epsilon^{-5/2}$, and $\epsilon^{-3/2}$ iterations (each iteration using $O(1)$ samples) when stochastic noises are in both level objectives, only in the upper-level objective, and not present (deterministic settings), respectively. We further show that if we employ momentum-assisted gradient estimators, the iteration complexities can be improved to $\epsilon^{-5/2}, \epsilon^{-4/2}$, and $\epsilon^{-3/2}$, respectively. We demonstrate even superior practical performance of the proposed method over existing second-order based approaches on MNIST data-hypercleaning experiments.
****

Complexity of Block Coordinate Descent with Proximal Regularization and Applications to Wasserstein CP-dictionary Learning
Dohyun Kwon, Hanbaek Lyu
We consider the block coordinate descent methods of Gauss-Seidel type with proximal regularization (BCD-PR), which is a classical method of minimizing general nonconvex objectives under constraints that has a wide range of practical applications. We theoretically establish the worst-case complexity bound for this algorithm. Namely, we show that for general nonconvex smooth objectives with block-wise constraints, the classical BCD-PR algorithm converges to an epsilon-stationary point within O(1/epsilon) iterations. Under a mild condition, this result still holds even if the algorithm is executed inexactly in each step. As an application, we propose a provable and efficient algorithm for ‘Wasserstein CP-dictionary learning’, which seeks a set of elementary probability distributions that can well-approximate a given set of d-dimensional joint probability distributions. Our algorithm is a version of BCD-PR that operates in the dual space, where the primal problem is regularized both entropically and proximally.
****

Data-OOB: Out-of-bag Estimate as a Simple and Efficient Data Value
Yongchan Kwon, James Zou
Data valuation is a powerful framework for providing statistical insights into w
hich data are beneficial or detrimental to model training. Many Shapley-based da
ta valuation methods have shown promising results in various downstream tasks, h
owever, they are well known to be computationally challenging as it requires tra
ining a large number of models. As a result, it has been recognized as infeasibl
e to apply to large datasets. To address this issue, we propose Data-OOB, a new
data valuation method for a bagging model that utilizes the out-of-bag estimate.
 The proposed method is computationally efficient and can scale to millions of d
ata by reusing trained weak learners. Specifically, Data-OOB takes less than $2.
25$ hours on a single CPU processor when there are 10^6 samples to evaluate an
d the input dimension is 100. Furthermore, Data-OOB has solid theoretical inte
rpretations in that it identifies the same important data point as the infinites
imal jackknife influence function when two different points are compared. We con
duct comprehensive experiments using 12 classification datasets, each with thous
ands of sample sizes. We demonstrate that the proposed method significantly outp
erforms existing state-of-the-art data valuation methods in identifying mislabel
ed data and finding a set of helpful (or harmful) data points, highlighting the
potential for applying data values in real-world applications.
****

Emergence of Adaptive Circadian Rhythms in Deep Reinforcement Learning
Aqeel Labash, Florian Stelzer, Daniel Majoral, Raul Vicente Zafra
Adapting to regularities of the environment is critical for biological organisms
 to anticipate events and plan. A prominent example is the circadian rhythm corr
esponding to the internalization by organisms of the 24-hour period of the Ear
th’s rotation. In this work, we study the emergence of circadian-like rhythms in
 deep reinforcement learning agents. In particular, we deployed agents in an env
ironment with a reliable periodic variation while solving a foraging task. We sy
stematically characterize the agent’s behavior during learning and demonstrate t
he emergence of a rhythm that is endogenous and entrainable. Interestingly, the
internal rhythm adapts to shifts in the phase of the environmental signal withou
t any re-training. Furthermore, we show via bifurcation and phase response curve
 analyses how artificial neurons develop dynamics to support the internalization
 of the environmental rhythm. From a dynamical systems view, we demonstrate that
 the adaptation proceeds by the emergence of a stable periodic orbit in the neur
on dynamics with a phase response that allows an optimal phase synchronisation b
etween the agent’s dynamics and the environmental rhythm.
****

Synergies between Disentanglement and Sparsity: Generalization and Identifiabili
ty in Multi-Task Learning
Sebastien Lachapelle, Tristan Deleu, Divyat Mahajan, Ioannis Mitliagkas, Yoshua
Bengio, Simon Lacoste-Julien, Quentin Bertrand
Although disentangled representations are often said to be beneficial for downst
ream tasks, current empirical and theoretical understanding is limited. In this
work, we provide evidence that disentangled representations coupled with sparse
task-specific predictors improve generalization. In the context of multi-task le
arning, we prove a new identifiability result that provides conditions under whi
ch maximally sparse predictors yield disentangled representations. Motivated by
this theoretical result, we propose a practical approach to learn disentangled r
epresentations based on a sparsity-promoting bi-level optimization problem. Fina
lly, we explore a meta-learning version of this algorithm based on group Lasso m
ulticlass SVM predictors, for which we derive a tractable dual formulation. It o
btains competitive results on standard few-shot classification benchmarks, while
 each task is using only a fraction of the learned representations.
****

Nearly-Optimal Hierarchical Clustering for Well-Clustered Graphs
Steinar Laenen, Bogdan Adrian Manghiuc, He Sun
This paper presents two efficient hierarchical clustering (HC) algorithms with r
espect to Dasgupta’s cost function. For any input graph G with a clear cluster

-structure, our designed algorithms run in nearly-linear time in the input size of G, and return an $O(1)$-approximate HC tree with respect to Dasgupta’s cost function. We compare the performance of our algorithm against the previous state-of-the-art on synthetic and real-world datasets and show that our designed algorithm produces comparable or better HC trees with much lower running time.
****

Hybrid Energy Based Model in the Feature Space for Out-of-Distribution Detection
Marc Lafon, Elias Ramzi, Clément Rambour, Nicolas Thome
Out-of-distribution (OOD) detection is a critical requirement for the deployment of deep neural networks. This paper introduces the HEAT model, a new post-hoc OOD detection method estimating the density of in-distribution (ID) samples using hybrid energy-based models (EBM) in the feature space of a pre-trained backbone. HEAT complements prior density estimators of the ID density, e.g. parametric models like the Gaussian Mixture Model (GMM), to provide an accurate yet robust density estimation. A second contribution is to leverage the EBM framework to provide a unified density estimation and to compose several energy terms. Extensive experiments demonstrate the significance of the two contributions. HEAT sets new state-of-the-art OOD detection results on the CIFAR-10 / CIFAR-100 benchmark as well as on the large-scale Imagenet benchmark. The code is available at: https://github.com/MarcLafon/heatood.
****

A theory of continuous generative flow networks
Salem Lahlou, Tristan Deleu, Pablo Lemos, Dinghuai Zhang, Alexandra Volokhova, Alex Hernández-Garc■■a, Lena Nehale Ezzine, Yoshua Bengio, Nikolay Malkin
Generative flow networks (GFlowNets) are amortized variational inference algorithms that are trained to sample from unnormalized target distributions over compositional objects. A key limitation of GFlowNets until this time has been that they are restricted to discrete spaces. We present a theory for generalized GFlowNets, which encompasses both existing discrete GFlowNets and ones with continuous or hybrid state spaces, and perform experiments with two goals in mind. First, we illustrate critical points of the theory and the importance of various assumptions. Second, we empirically demonstrate how observations about discrete GFlowNets transfer to the continuous case and show strong results compared to non-GFlowNet baselines on several previously studied tasks. This work greatly widens the perspectives for the application of GFlowNets in probabilistic inference and various modeling settings.
****

Automatically marginalized MCMC in probabilistic programming
Jinlin Lai, Javier Burroni, Hui Guan, Daniel Sheldon
Hamiltonian Monte Carlo (HMC) is a powerful algorithm to sample latent variables from Bayesian models. The advent of probabilistic programming languages (PPLs) frees users from writing inference algorithms and lets users focus on modeling. However, many models are difficult for HMC to solve directly, and often require tricks like model reparameterization. We are motivated by the fact that many of those models could be simplified by marginalization. We propose to use automatic marginalization as part of the sampling process using HMC in a graphical model extracted from a PPL, which substantially improves sampling from real-world hierarchical models.
****

DS-1000: A Natural and Reliable Benchmark for Data Science Code Generation
Yuhang Lai, Chengxi Li, Yiming Wang, Tianyi Zhang, Ruiqi Zhong, Luke Zettlemoyer, Wen-Tau Yih, Daniel Fried, Sida Wang, Tao Yu
We introduce DS-1000, a code generation benchmark with a thousand data science problems spanning seven Python libraries, such as Numpy and Pandas. Compared to prior works, DS-1000 incorporates three core features. First, our problems reflect diverse, realistic, and practical use cases since we collected them from Stack Overflow. Second, our automatic evaluation is highly specific (reliable) – across all Codex-002-predicted solutions that our evaluation accepts, only 1.8% of them are incorrect; we achieve this with multi-criteria metrics, checking both functional correctness by running test cases and surface-form constraints by restri

cting API usages or keywords. Finally, we proactively defend against memorizatio
n by slightly modifying our problems to be different from the original StackOver
flow source; consequently, models cannot answer them correctly by memorizing the
 solutions from pre-training. The current best public system (Codex-002) achieve
s 43.3% accuracy, leaving ample room for improvement. We release our benchmark a
t https://ds1000-code-gen.github.io.
****

ChiPFormer: Transferable Chip Placement via Offline Decision Transformer
Yao Lai, Jinxin Liu, Zhentao Tang, Bin Wang, Jianye Hao, Ping Luo
Placement is a critical step in modern chip design, aiming to determine the posi
tions of circuit modules on the chip canvas. Recent works have shown that reinfo
rcement learning (RL) can improve human performance in chip placement. However,
 such an RL-based approach suffers from long training time and low transfer abili
ty in unseen chip circuits. To resolve these challenges, we cast the chip placem
ent as an offline RL formulation and present ChiPFormer that enables learning a
 transferable placement policy from fixed offline data. ChiPFormer has several ad
vantages that prior arts do not have. First, ChiPFormer can exploit offline plac
ement designs to learn transferable policies more efficiently in a multi-task se
tting. Second, ChiPFormer can promote effective finetuning for unseen chip circu
its, reducing the placement runtime from hours to minutes. Third, extensive expe
riments on 32 chip circuits demonstrate that ChiPFormer achieves significantly b
etter placement quality while reducing the runtime by 10x compared to recent sta
te-of-the-art approaches in both public benchmarks and realistic industrial task
s. The deliverables are released at https://sites.google.com/view/chipformer/hom
e.
****

FP-Diffusion: Improving Score-based Diffusion Models by Enforcing the Underlying
 Score Fokker-Planck Equation
Chieh-Hsin Lai, Yuhta Takida, Naoki Murata, Toshimitsu Uesaka, Yuki Mitsufuji, S
tefano Ermon
Score-based generative models (SGMs) learn a family of noise-conditional score f
unctions corresponding to the data density perturbed with increasingly large amo
unts of noise. These perturbed data densities are linked together by the Fokker-
Planck equation (FPE), a partial differential equation (PDE) governing the spati
al-temporal evolution of a density undergoing a diffusion process. In this work,
 we derive a corresponding equation called the score FPE that characterizes the
noise-conditional scores of the perturbed data densities (i.e., their gradients)
. Surprisingly, despite the impressive empirical performance, we observe that sc
ores learned through denoising score matching (DSM) fail to fulfill the underlyi
ng score FPE, which is an inherent self-consistency property of the ground truth
 score. We prove that satisfying the score FPE is desirable as it improves the l
ikelihood and the degree of conservativity. Hence, we propose to regularize the
DSM objective to enforce satisfaction of the score FPE, and we show the effectiv
eness of this approach across various datasets.
****

Private Statistical Estimation of Many Quantiles
Clément Lalanne, Aurélien Garivier, Rémi Gribonval
This work studies the estimation of many statistical quantiles under differentia
l privacy. More precisely, given a distribution and access to i.i.d. samples fro
m it, we study the estimation of the inverse of its cumulative distribution func
tion (the quantile function) at specific points. For instance, this task is of k
ey importance in private data generation. We present two different approaches. T
he first one consists in privately estimating the empirical quantiles of the sam
ples and using this result as an estimator of the quantiles of the distribution.
 In particular, we study the statistical properties of the recently published al
gorithm introduced by (Kaplan et al., 2022) that privately estimates the quantil
es recursively. The second approach is to use techniques of density estimation i
n order to uniformly estimate the quantile function on an interval. In particula
r, we show that there is a tradeoff between the two methods. When we want to est
imate many quantiles, it is better to estimate the density rather than estimatin

g the quantile function at specific points.
****

Bootstrap in High Dimension with Low Computation
Henry Lam, Zhenyuan Liu
The bootstrap is a popular data-driven method to quantify statistical uncertainty, but for modern high-dimensional problems, it could suffer from huge computational costs due to the need to repeatedly generate resamples and refit models. We study the use of bootstraps in high-dimensional environments with a small number of resamples. In particular, we show that with a recent "cheap" bootstrap perspective, using a number of resamples as small as one could attain valid coverage even when the dimension grows closely with the sample size, thus strongly supporting the implementability of the bootstrap for large-scale problems. We validate our theoretical results and compare the performance of our approach with other benchmarks via a range of experiments.
****

LegendreTron: Uprising Proper Multiclass Loss Learning
Kevin H Lam, Christian Walder, Spiridon Penev, Richard Nock
Loss functions serve as the foundation of supervised learning and are often chosen prior to model development. To avoid potentially ad hoc choices of losses, statistical decision theory describes a desirable property for losses known as properness, which asserts that Bayes' rule is optimal. Recent works have sought to learn losses and models jointly. Existing methods do this by fitting an inverse canonical link function which monotonically maps $\mathbb{R}$ to $[0,1]$ to estimate probabilities for binary problems. In this paper, we extend monotonicity to maps between $\mathbb{R}^{C-1}$ and the projected probability simplex $\tilde{\Delta}^{C-1}$ by using monotonicity of gradients of convex functions. We present LegendreTron as a novel and practical method that jointly learns proper canonical losses and probabilities for multiclass problems. Tested on a benchmark of domains with up to 1,000 classes, our experimental results show that our method consistently outperforms the natural multiclass baseline under a t-test at 99% significance on all datasets with greater than 10 classes.
****

Metagenomic Binning using Connectivity-constrained Variational Autoencoders
Andre Lamurias, Alessandro Tibo, Katja Hose, Mads Albertsen, Thomas Dyhre Nielsen
Current state-of-the-art techniques for metagenomic binning only utilize local features for the individual DNA sequences (contigs), neglecting additional information such as the assembly graph, in which the contigs are connected according to overlapping reads, and gene markers identified in the contigs. In this paper, we propose the use of a Variational AutoEncoder (VAE) tailored to leverage auxiliary structural information about contig relations when learning contig representations for subsequent metagenomic binning. Our method, CCVAE, improves on previous work that used VAEs for learning latent representations of the individual contigs, by constraining these representations according to the connectivity information from the assembly graph. Additionally, we incorporate into the model additional information in the form of marker genes to better differentiate contigs from different genomes. Our experiments on both simulated and real-world datasets demonstrate that CCVAE outperforms current state-of-the-art techniques, thus providing a more effective method for metagenomic binning.
****

Delay-Adapted Policy Optimization and Improved Regret for Adversarial MDP with Delayed Bandit Feedback
Tal Lancewicki, Aviv Rosenberg, Dmitry Sotnikov
Policy Optimization (PO) is one of the most popular methods in Reinforcement Learning (RL). Thus, theoretical guarantees for PO algorithms have become especially important to the RL community. In this paper, we study PO in adversarial MDPs with a challenge that arises in almost every real-world application – delayed bandit feedback. We give the first near-optimal regret bounds for PO in tabular MDPs, and may even surpass state-of-the-art (which uses less efficient methods). Our novel Delay-Adapted PO (DAPO) is easy to implement and to generalize, allowin

g us to extend our algorithm to: (i) infinite state space under the assumption o
f linear Q-function, proving the first regret bounds for delayed feedback with
 function approximation. (ii) deep RL, demonstrating its effectiveness in experi
ments on MuJoCo domains.
****

Lottery Tickets in Evolutionary Optimization: On Sparse Backpropagation-Free Tra
inability

Robert Tjarko Lange, Henning Sprekeler

Is the lottery ticket phenomenon an idiosyncrasy of gradient-based training or d
oes it generalize to evolutionary optimization? In this paper we establish the e
xistence of highly sparse trainable initializations for evolution strategies (ES
) and characterize qualitative differences compared to gradient descent (GD)-bas
ed sparse training. We introduce a novel signal-to-noise iterative pruning proce
dure, which incorporates loss curvature information into the network pruning ste
p. This can enable the discovery of even sparser trainable network initializatio
ns when using black-box evolution as compared to GD-based optimization. Furtherm
ore, we find that these initializations encode an inductive bias, which transfer
s across different ES, related tasks and even to GD-based training. Finally, we
compare the local optima resulting from the different optimization paradigms and
 sparsity levels. In contrast to GD, ES explore diverse and flat local optima an
d do not preserve linear mode connectivity across sparsity levels and independen
t runs. The results highlight qualitative differences between evolution and grad
ient-based learning dynamics, which can be uncovered by the study of iterative p
runing procedures.
****

On the Occupancy Measure of Non-Markovian Policies in Continuous MDPs

Romain Laroche, Remi Tachet Des Combes

The state-action occupancy measure of a policy is the expected (discounted or un
discounted) number of times a state-action couple is visited in a trajectory. Fo
r decades, RL books have been reporting the occupancy equivalence between Markov
ian and non-Markovian policies in countable state-action spaces under mild condi
tions. This equivalence states that the occupancy of any non-Markovian policy ca
n be equivalently obtained by a Markovian policy, i.e. a memoryless probability
distribution, conditioned only on its current state. While expected, for technic
al reasons, the translation of this result to continuous state space has resiste
d until now. Our main contribution is to fill this gap and to provide a general
measure-theoretic treatment of the problem, permitting, in particular, its exten
sion to continuous MDPs. Furthermore, we show that when the occupancy is infinit
e, we may encounter some non-trivial cases where the result does not hold anymor
e.
****

Minimalistic Predictions to Schedule Jobs with Online Precedence Constraints

Alexandra Anna Lassota, Alexander Lindermayr, Nicole Megow, Jens Schlöter

We consider non-clairvoyant scheduling with online precedence constraints, where
 an algorithm is oblivious to any job dependencies and learns about a job only i
f all of its predecessors have been completed. Given strong impossibility result
s in classical competitive analysis, we investigate the problem in a learning-au
gmented setting, where an algorithm has access to predictions without any qualit
y guarantee. We discuss different prediction models: novel problem-specific mode
ls as well as general ones, which have been proposed in previous works. We prese
nt lower bounds and algorithmic upper bounds for different precedence topologies
, and thereby give a structured overview on which and how additional (possibly e
rroneous) information helps for designing better algorithms. Along the way, we a
lso improve bounds on traditional competitive ratios for existing algorithms.
****

Speeding Up Bellman Ford via Minimum Violation Permutations

Silvio Lattanzi, Ola Svensson, Sergei Vassilvitskii

The Bellman-Ford algorithm is a basic primitive for computing single source shor
test paths in graphs with negative weight edges. Its running time is governed by
 the order the algorithm examines vertices for iterative updates on the value of

their shortest path. In this work we study this problem through the lens of ’Algorithms with predictions,’ and show how to leverage auxiliary information from similar instances to improve the running time. We do this by identifying the key problem of Minimum Violation Permutations, and give algorithms with strong approximation guarantees as well as formal lower bounds. We complement the theoretical analysis with an empirical evaluation, showing that this approach can lead to a significant speed up in practice.

****

Who Needs to Know? Minimal Knowledge for Optimal Coordination

Niklas Lauffer, Ameesh Shah, Micah Carroll, Michael D Dennis, Stuart Russell

To optimally coordinate with others in cooperative games, it is often crucial to have information about one’s collaborators: successful driving requires understanding which side of the road to drive on. However, not every feature of collaborators is strategically relevant: the fine-grained acceleration of drivers may be ignored while maintaining optimal coordination. We show that there is a well-defined dichotomy between strategically relevant and irrelevant information. Moreover, we show that, in dynamic games, this dichotomy has a compact representation that can be efficiently computed via a Bellman backup operator. We apply this algorithm to analyze the strategically relevant information for tasks in both a standard and a partially observable version of the Overcooked environment. Theoretical and empirical results show that our algorithms are significantly more efficient than baselines. Videos are available at https://minknowledge.github.io.

****

Target-based Surrogates for Stochastic Optimization

Jonathan Wilder Lavington, Sharan Vaswani, Reza Babanezhad Harikandeh, Mark Schmidt, Nicolas Le Roux

We consider minimizing functions for which it is expensive to compute the (possibly stochastic) gradient. Such functions are prevalent in reinforcement learning, imitation learning and adversarial training. Our target optimization framework uses the (expensive) gradient computation to construct surrogate functions in a target space (e.g. the logits output by a linear model for classification) that can be minimized efficiently. This allows for multiple parameter updates to the model, amortizing the cost of gradient computation. In the full-batch setting, we prove that our surrogate is a global upper-bound on the loss, and can be (locally) minimized using a black-box optimization algorithm. We prove that the resulting majorization-minimization algorithm ensures convergence to a stationary point of the loss. Next, we instantiate our framework in the stochastic setting and propose the SSO algorithm, which can be viewed as projected stochastic gradient descent in the target space. This connection enables us to prove theoretical guarantees for SSO when minimizing convex functions. Our framework allows the use of standard stochastic optimization algorithms to construct surrogates which can be minimized by any deterministic optimization method. To evaluate our framework, we consider a suite of supervised learning and imitation learning problems. Our experiments indicate the benefits of target optimization and the effectiveness of SSO.

****

Cluster Explanation via Polyhedral Descriptions

Connor Lawless, Oktay Gunluk

This paper focuses on the cluster description problem where, given a dataset and its partition into clusters, the task is to explain the clusters. We introduce a new approach to explain clusters by constructing a polyhedron around each cluster while minimizing either the complexity of the resulting polyhedra or the number of features used in the description. We formulate the cluster description problem as an integer program and present a column generation approach to search over an exponential number of candidate half-spaces that can be used to build the polyhedra. To deal with large datasets, we introduce a novel grouping scheme that at first forms smaller groups of data points and then builds the polyhedra around the grouped data, a strategy which out-performs the common approach of sub-sampling data. Compared to state of the art cluster description algorithms, our approach is able to achieve competitive interpretability with improved description

accuracy.
****
Pre-training for Speech Translation: CTC Meets Optimal Transport
Phuong-Hang Le, Hongyu Gong, Changhan Wang, Juan Pino, Benjamin Lecouteux, Didier Schwab
The gap between speech and text modalities is a major challenge in speech-to-text translation (ST). Different methods have been proposed to reduce this gap, but most of them require architectural changes in ST training. In this work, we propose to mitigate this issue at the pre-training stage, requiring no change in the ST model. First, we show that the connectionist temporal classification (CTC) loss can reduce the modality gap by design. We provide a quantitative comparison with the more common cross-entropy loss, showing that pre-training with CTC consistently achieves better final ST accuracy. Nevertheless, CTC is only a partial solution and thus, in our second contribution, we propose a novel pre-training method combining CTC and optimal transport to further reduce this gap. Our method pre-trains a Siamese-like model composed of two encoders, one for acoustic inputs and the other for textual inputs, such that they produce representations that are close to each other in the Wasserstein space. Extensive experiments on the standard CoVoST-2 and MuST-C datasets show that our pre-training method applied to the vanilla encoder-decoder Transformer achieves state-of-the-art performance under the no-external-data setting, and performs on par with recent strong multi-task learning systems trained with external data. Finally, our method can also be applied on top of these multi-task systems, leading to further improvements for these models.
****
Bootstrapped Representations in Reinforcement Learning
Charline Le Lan, Stephen Tu, Mark Rowland, Anna Harutyunyan, Rishabh Agarwal, Marc G Bellemare, Will Dabney
In reinforcement learning (RL), state representations are key to dealing with large or continuous state spaces. While one of the promises of deep learning algorithms is to automatically construct features well-tuned for the task they try to solve, such a representation might not emerge from end-to-end training of deep RL agents. To mitigate this issue, auxiliary objectives are often incorporated into the learning process and help shape the learnt state representation. Bootstrapping methods are today’s method of choice to make these additional predictions. Yet, it is unclear which features these algorithms capture and how they relate to those from other auxiliary-task-based approaches. In this paper, we address this gap and provide a theoretical characterization of the state representation learnt by temporal difference learning (Sutton, 1988). Surprisingly, we find that this representation differs from the features learned by Monte Carlo and residual gradient algorithms for most transition structures of the environment in the policy evaluation setting. We describe the efficacy of these representations for policy evaluation, and use our theoretical analysis to design new auxiliary learning rules. We complement our theoretical results with an empirical comparison of these learning rules for different cumulant functions on classic domains such as the four-room domain (Sutton et al, 1999) and Mountain Car (Moore, 1990).
****
Strategic Classification with Unknown User Manipulations
Tosca Lechner, Ruth Urner, Shai Ben-David
In many human-centric applications for Machine Learning instances will adapt to a classifier after its deployment. The field of strategic classification deals with this issue by aiming for a classifier that balances the trade-off between correctness and robustness to manipulation. This task is made harder if the underlying manipulation structure (i.e. the set of manipulations available at every instance) is unknown to the learner. We propose a novel batch-learning setting in which we use unlabeled data from previous rounds to estimate the manipulation structure. We show that in this batch-learning setting it is possible to learn a close-to-optimal classifier in terms of the strategic loss even without knowing the feasible manipulations beforehand. In line with recent advances in the strategic classification literature, we do not assume a best-response from agents but

only require that observed manipulations are feasible.
****
Learning in POMDPs is Sample-Efficient with Hindsight Observability
Jonathan Lee, Alekh Agarwal, Christoph Dann, Tong Zhang
POMDPs capture a broad class of decision making problems, but hardness results suggest that learning is intractable even in simple settings due to the inherent partial observability. However, in many realistic problems, more information is either revealed or can be computed during some point of the learning process. Motivated by diverse applications ranging from robotics to data center scheduling, we formulate a Hindsight Observable Markov Decision Process (HOMDP) as a POMDP where the latent states are revealed to the learner in hindsight and only during training. We introduce new algorithms for the tabular and function approximation settings that are provably sample-efficient with hindsight observability, even in POMDPs that would otherwise be statistically intractable. We give a lower bound showing that the tabular algorithm is optimal in its dependence on latent state and observation cardinalities.
****
Towards Deep Attention in Graph Neural Networks: Problems and Remedies
Soo Yong Lee, Fanchen Bu, Jaemin Yoo, Kijung Shin
Graph neural networks (GNNs) learn the representation of graph-structured data, and their expressiveness can be further enhanced by inferring node relations for propagation. Attention-based GNNs infer neighbor importance to manipulate the weight of its propagation. Despite their popularity, the discussion on deep graph attention and its unique challenges has been limited. In this work, we investigate some problematic phenomena related to deep graph attention, including vulnerability to over-smoothed features and smooth cumulative attention. Through theoretical and empirical analyses, we show that various attention-based GNNs suffer from these problems. Motivated by our findings, we propose AERO-GNN, a novel GNN architecture designed for deep graph attention. AERO-GNN provably mitigates the proposed problems of deep graph attention, which is further empirically demonstrated with (a) its adaptive and less smooth attention functions and (b) higher performance at deep layers (up to 64). On 9 out of 12 node classification benchmarks, AERO-GNN outperforms the baseline GNNs, highlighting the advantages of deep graph attention. Our code is available at https://github.com/syleeheal/AERO-GNN.
****
InGram: Inductive Knowledge Graph Embedding via Relation Graphs
Jaejun Lee, Chanyoung Chung, Joyce Jiyoung Whang
Inductive knowledge graph completion has been considered as the task of predicting missing triplets between new entities that are not observed during training. While most inductive knowledge graph completion methods assume that all entities can be new, they do not allow new relations to appear at inference time. This restriction prohibits the existing methods from appropriately handling real-world knowledge graphs where new entities accompany new relations. In this paper, we propose an INductive knowledge GRAph eMbedding method, InGram, that can generate embeddings of new relations as well as new entities at inference time. Given a knowledge graph, we define a relation graph as a weighted graph consisting of relations and the affinity weights between them. Based on the relation graph and the original knowledge graph, InGram learns how to aggregate neighboring embeddings to generate relation and entity embeddings using an attention mechanism. Experimental results show that InGram outperforms 14 different state-of-the-art methods on varied inductive learning scenarios.
****
Optimality of Thompson Sampling with Noninformative Priors for Pareto Bandits
Jongyeong Lee, Junya Honda, Chao-Kai Chiang, Masashi Sugiyama
In the stochastic multi-armed bandit problem, a randomized probability matching policy called Thompson sampling (TS) has shown excellent performance in various reward models. In addition to the empirical performance, TS has been shown to achieve asymptotic problem-dependent lower bounds in several models. However, its optimality has been mainly addressed under light-tailed or one-parameter models

that belong to exponential families. In this paper, we consider the optimality o
f TS for the Pareto model that has a heavy tail and is parameterized by two unkn
own parameters. Specifically, we discuss the optimality of TS with probability m
atching priors that include the Jeffreys prior and the reference priors. We firs
t prove that TS with certain probability matching priors can achieve the optimal
 regret bound. Then, we show the suboptimality of TS with other priors, includin
g the Jeffreys and the reference priors. Nevertheless, we find that TS with the
Jeffreys and reference priors can achieve the asymptotic lower bound if one uses
 a truncation procedure. These results suggest carefully choosing noninformative
 priors to avoid suboptimality and show the effectiveness of truncation procedur
es in TS-based policies.
****

Conditional Graph Information Bottleneck for Molecular Relational Learning
Namkyeong Lee, Dongmin Hyun, Gyoung S. Na, Sungwon Kim, Junseok Lee, Chanyoung P
ark
Molecular relational learning, whose goal is to learn the interaction behavior b
etween molecular pairs, got a surge of interest in molecular sciences due to its
 wide range of applications. Recently, graph neural networks have recently shown
 great success in molecular relational learning by modeling a molecule as a grap
h structure, and considering atom-level interactions between two molecules. Desp
ite their success, existing molecular relational learning methods tend to overlo
ok the nature of chemistry, i.e., a chemical compound is composed of multiple su
bstructures such as functional groups that cause distinctive chemical reactions.
 In this work, we propose a novel relational learning framework, called CGIB, th
at predicts the interaction behavior between a pair of graphs by detecting core
subgraphs therein. The main idea is, given a pair of graphs, to find a subgraph
from a graph that contains the minimal sufficient information regarding the task
 at hand conditioned on the paired graph based on the principle of conditional g
raph information bottleneck. We argue that our proposed method mimics the nature
 of chemical reactions, i.e., the core substructure of a molecule varies dependi
ng on which other molecule it interacts with. Extensive experiments on various t
asks with real-world datasets demonstrate the superiority of CGIB over state-of-
the-art baselines. Our code is available at https://github.com/Namkyeong/CGIB.
****

Exploring Chemical Space with Score-based Out-of-distribution Generation
Seul Lee, Jaehyeong Jo, Sung Ju Hwang
A well-known limitation of existing molecular generative models is that the gene
rated molecules highly resemble those in the training set. To generate truly nov
el molecules that may have even better properties for de novo drug discovery, mo
re powerful exploration in the chemical space is necessary. To this end, we prop
ose Molecular Out-Of-distribution Diffusion(MOOD), a score-based diffusion schem
e that incorporates out-of-distribution (OOD) control in the generative stochast
ic differential equation (SDE) with simple control of a hyperparameter, thus req
uires no additional costs. Since some novel molecules may not meet the basic req
uirements of real-world drugs, MOOD performs conditional generation by utilizing
 the gradients from a property predictor that guides the reverse-time diffusion
process to high-scoring regions according to target properties such as protein-l
igand interactions, drug-likeness, and synthesizability. This allows MOOD to sea
rch for novel and meaningful molecules rather than generating unseen yet trivial
 ones. We experimentally validate that MOOD is able to explore the chemical spac
e beyond the training distribution, generating molecules that outscore ones foun
d with existing methods, and even the top 0.01% of the original training pool. O
ur code is available at https://github.com/SeulLee05/MOOD.
****

Pix2Struct: Screenshot Parsing as Pretraining for Visual Language Understanding
Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Mart
in Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, Kristina Toutano
va
Visually-situated language is ubiquitous—sources range from textbooks with diagr
ams to web pages with images and tables, to mobile apps with buttons and forms.

Perhaps due to this diversity, previous work has typically relied on domain-spec
ific recipes with limited sharing of the underlying data, model architectures, a
nd objectives. We present Pix2Struct, a pretrained image-to-text model for purel
y visual language understanding, which can be finetuned on tasks containing visu
ally-situated language. Pix2Struct is pretrained by learning to parse masked scr
eenshots of web pages into simplified HTML. The web, with its richness of visual
 elements cleanly reflected in the HTML structure, provides a large source of pr
etraining data well suited to the diversity of downstream tasks. Intuitively, th
is objective subsumes common pretraining signals such as OCR, language modeling,
 and image captioning. In addition to the novel pretraining strategy, we introdu
ce a variable-resolution input representation and a more flexible integration of
 language and vision inputs, where language prompts such as questions are render
ed directly on top of the input image. For the first time, we show that a single
 pretrained model can achieve state-of-the-art results in six out of nine tasks
across four domains: documents, illustrations, user interfaces, and natural imag
es.
****

FlexRound: Learnable Rounding based on Element-wise Division for Post-Training Q
uantization
Jung Hyun Lee, Jeonghoon Kim, Se Jung Kwon, Dongsoo Lee
Post-training quantization (PTQ) has been gaining popularity for the deployment
of deep neural networks on resource-limited devices since unlike quantization-aw
are training, neither a full training dataset nor end-to-end training is require
d at all. As PTQ schemes based on reconstructing each layer or block output turn
 out to be effective to enhance quantized model performance, recent works have d
eveloped algorithms to devise and learn a new weight-rounding scheme so as to be
tter reconstruct each layer or block output. In this work, we propose a simple y
et effective new weight-rounding mechanism for PTQ, coined FlexRound, based on e
lement-wise division instead of typical element-wise addition such that FlexRoun
d enables jointly learning a common quantization grid size as well as a differen
t scale for each pre-trained weight. Thanks to the reciprocal rule of derivative
s induced by element-wise division, FlexRound is inherently able to exploit pre-
trained weights when updating their corresponding scales, and thus, flexibly qua
ntize pre-trained weights depending on their magnitudes. We empirically validate
 the efficacy of FlexRound on a wide range of models and tasks. To the best of o
ur knowledge, our work is the first to carry out comprehensive experiments on no
t only image classification and natural language understanding but also natural
language generation, assuming a per-tensor uniform PTQ setting. Moreover, we dem
onstrate, for the first time, that large language models can be efficiently quan
tized, with only a negligible impact on performance compared to half-precision b
aselines, achieved by reconstructing the output in a block-by-block manner.
****

CoDi: Co-evolving Contrastive Diffusion Models for Mixed-type Tabular Synthesis
Chaejeong Lee, Jayoung Kim, Noseong Park
With growing attention to tabular data these days, the attempt to apply a synthe
tic table to various tasks has been expanded toward various scenarios. Owing to
the recent advances in generative modeling, fake data generated by tabular data
synthesis models become sophisticated and realistic. However, there still exists
 a difficulty in modeling discrete variables (columns) of tabular data. In this
work, we propose to process continuous and discrete variables separately (but be
ing conditioned on each other) by two diffusion models. The two diffusion models
 are co-evolved during training by reading conditions from each other. In order
to further bind the diffusion models, moreover, we introduce a contrastive learn
ing method with a negative sampling method. In our experiments with 11 real-worl
d tabular datasets and 8 baseline methods, we prove the efficacy of the proposed
 method, called $\texttt{CoDi}$. Our code is available at https://github.com/Cha
ejeongLee/CoDi.
****

Minimizing Trajectory Curvature of ODE-based Generative Models
Sangyun Lee, Beomsu Kim, Jong Chul Ye

Recent ODE/SDE-based generative models, such as diffusion models, rectified flow
s, and flow matching, define a generative process as a time reversal of a fixed
forward process. Even though these models show impressive performance on large-s
cale datasets, numerical simulation requires multiple evaluations of a neural ne
twork, leading to a slow sampling speed. We attribute the reason to the high cur
vature of the learned generative trajectories, as it is directly related to the
truncation error of a numerical solver. Based on the relationship between the fo
rward process and the curvature, here we present an efficient method of training
 the forward process to minimize the curvature of generative trajectories withou
t any ODE/SDE simulation. Experiments show that our method achieves a lower curv
ature than previous models and, therefore, decreased sampling costs while mainta
ining competitive performance. Code is available at https://github.com/sangyun88
4/fast-ode.
****

H-Likelihood Approach to Deep Neural Networks with Temporal-Spatial Random Effec
ts for High-Cardinality Categorical Features

Hangbin Lee, Youngjo Lee

Deep Neural Networks (DNNs) are one of the most powerful tools for prediction, b
ut many of them implicitly assume that the data are statistically independent. H
owever, in the real world, it is common for large-scale data to be clustered wit
h temporal-spatial correlation structures. Variational approaches and integrated
 likelihood approaches have been proposed to obtain approximate maximum likeliho
od estimators (MLEs) for correlated data. However, due to the large size of data
, they cannot provide exact MLEs. In this study, we propose a new hierarchical l
ikelihood approach to DNNs with correlated random effects for clustered data. By
 jointly optimizing the the negative h-likelihood loss, we can provide exact MLE
s for both mean and dispersion parameters, as well as the best linear unbiased p
redictors for the random effects. Moreover, the hierarchical likelihood allows a
 computable procedure for restricted maximum likelihood estimators of dispersion
 parameters. The proposed two-step algorithm enables online learning for the neu
ral networks, whereas the integrated likelihood cannot decompose like a widely-u
sed loss function in DNNs. The proposed h-likelihood approach offers several adv
antages, which we demonstrate through numerical studies and real data analyses.
****

On the Importance of Feature Decorrelation for Unsupervised Representation Learn
ing in Reinforcement Learning

Hojoon Lee, Koanho Lee, Dongyoon Hwang, Hyunho Lee, Byungkun Lee, Jaegul Choo

Recently, unsupervised representation learning (URL) has improved the sample eff
iciency of Reinforcement Learning (RL) by pretraining a model from a large unlab
eled dataset. The underlying principle of these methods is to learn temporally p
redictive representations by predicting future states in the latent space. Howev
er, an important challenge of this approach is the representational collapse, wh
ere the subspace of the latent representations collapses into a low-dimensional
manifold. To address this issue, we propose a novel URL framework that causally
predicts future states while increasing the dimension of the latent manifold by
decorrelating the features in the latent space. Through extensive empirical stud
ies, we demonstrate that our framework effectively learns predictive representat
ions without collapse, which significantly improves the sample efficiency of sta
te-of-the-art URL methods on the Atari 100k benchmark. The code is available at
https://github.com/dojeon-ai/SimTPR.
****

HETAL: Efficient Privacy-preserving Transfer Learning with Homomorphic Encryptio
n

Seewoo Lee, Garam Lee, Jung Woo Kim, Junbum Shin, Mun-Kyu Lee

Transfer learning is a de facto standard method for efficiently training machine
 learning models for data-scarce problems by adding and fine-tuning new classifi
cation layers to a model pre-trained on large datasets. Although numerous previo
us studies proposed to use homomorphic encryption to resolve the data privacy is
sue in transfer learning in the machine learning as a service setting, most of t
hem only focused on encrypted inference. In this study, we present HETAL, an eff

icient Homomorphic Encryption based Transfer Learning algorithm, that protects t he client’s privacy in training tasks by encrypting the client data using the CK KS homomorphic encryption scheme. HETAL is the first practical scheme that stric tly provides encrypted training, adopting validation-based early stopping and ac hieving the accuracy of nonencrypted training. We propose an efficient encrypted matrix multiplication algorithm, which is 1.8 to 323 times faster than prior me thods, and a highly precise softmax approximation algorithm with increased cover age. The experimental results for five well-known benchmark datasets show total training times of 567–3442 seconds, which is less than an hour.
****

QASA: Advanced Question Answering on Scientific Articles
Yoonjoo Lee, Kyungjae Lee, Sunghyun Park, Dasol Hwang, Jaehyeon Kim, Hong-In Lee , Moontae Lee
Reasoning is the crux of intellectual thinking. While question answering (QA) ta sks are prolific with various computational models and benchmark datasets, they mostly tackle factoid or shallow QA without asking deeper understanding. Dual pr ocess theory asserts that human reasoning consists of associative thinking to co llect relevant pieces of knowledge and logical reasoning to consciously conclude grounding on evidential rationale. Based on our intensive think-aloud study tha t revealed the three types of questions: surface, testing, and deep questions, w e first propose the QASA benchmark that consists of 1798 novel question answerin g pairs that require full-stack reasoning on scientific articles in AI and ML fi elds. Then we propose the QASA approach that tackles the full-stack reasoning wi th large language models via associative selection, evidential rationale-generat ion, and systematic composition. Our experimental results show that QASA’s full- stack inference outperforms the state-of-the-art InstructGPT by a big margin. We also find that rationale-generation is critical for the performance gain, claim ing how we should rethink advanced question answering. The dataset is available at https://github.com/lgresearch/QASA.
****

Demystifying Disagreement-on-the-Line in High Dimensions
Donghwan Lee, Behrad Moniri, Xinmeng Huang, Edgar Dobriban, Hamed Hassani
Evaluating the performance of machine learning models under distribution shifts is challenging, especially when we only have unlabeled data from the shifted (ta rget) domain, along with labeled data from the original (source) domain. Recent work suggests that the notion of disagreement, the degree to which two models tr ained with different randomness differ on the same input, is a key to tackling t his problem. Experimentally, disagreement and prediction error have been shown t o be strongly connected, which has been used to estimate model performance. Expe riments have led to the discovery of the disagreement-on-the-line phenomenon, wh ereby the classification error under the target domain is often a linear functio n of the classification error under the source domain; and whenever this propert y holds, disagreement under the source and target domain follow the same linear relation. In this work, we develop a theoretical foundation for analyzing disagr eement in high-dimensional random features regression; and study under what cond itions the disagreement-on-the-line phenomenon occurs in our setting. Experiment s on CIFAR-10-C, Tiny ImageNet-C, and Camelyon17 are consistent with our theory and support the universality of the theoretical findings.
****

On the Correctness of Automatic Differentiation for Neural Networks with Machine -Representable Parameters
Wonyeol Lee, Sejun Park, Alex Aiken
Recent work has shown that forward- and reverse- mode automatic differentiation (AD) over the reals is almost always correct in a mathematically precise sense. However, actual programs work with machine-representable numbers (e.g., floating -point numbers), not reals. In this paper, we study the correctness of AD when t he parameter space of a neural network consists solely of machine-representable numbers. In particular, we analyze two sets of parameters on which AD can be inc orrect: the incorrect set on which the network is differentiable but AD does not compute its derivative, and the non-differentiable set on which the network is

non-differentiable. For a neural network with bias parameters, we first prove th
at the incorrect set is always empty. We then prove a tight bound on the size of
 the non-differentiable set, which is linear in the number of non-differentiabil
ities in activation functions, and give a simple necessary and sufficient condit
ion for a parameter to be in this set. We further prove that AD always computes
 a Clarke subderivative even on the non-differentiable set. We also extend these
 results to neural networks possibly without bias parameters.
****

Implicit Jacobian regularization weighted with impurity of probability output
Sungyoon Lee, Jinseong Park, Jaewook Lee
The success of deep learning is greatly attributed to stochastic gradient descen
t (SGD), yet it remains unclear how SGD finds well-generalized models. We demons
trate that SGD has an implicit regularization effect on the logit-weight Jacobia
n norm of neural networks. This regularization effect is weighted with the impur
ity of the probability output, and thus it is active in a certain phase of train
ing. Moreover, based on these findings, we propose a novel optimization method t
hat explicitly regularizes the Jacobian norm, which leads to similar performance
 as other state-of-the-art sharpness-aware optimization methods.
****

Unsupervised Skill Discovery for Learning Shared Structures across Changing Envi
ronments
Sang-Hyun Lee, Seung-Woo Seo
Learning shared structures across changing environments enables an agent to effi
ciently retain obtained knowledge and transfer it between environments. A skill
 is a promising concept to represent shared structures. Several recent works prop
osed unsupervised skill discovery algorithms that can discover useful skills wit
hout a reward function. However, they focused on discovering skills in stationar
y environments or assumed that a skill being trained is fixed within an episode,
 which is insufficient to learn and represent shared structures. In this paper,
we introduce a new unsupervised skill discovery algorithm that discovers a set o
f skills that can represent shared structures across changing environments. Our
algorithm trains incremental skills and encourages a new skill to expand state c
overage obtained with compositions of previously learned skills. We also introduc
e a skill evaluation process to prevent our skills from containing redundant sk
ills, a common issue in previous work. Our experimental results show that our al
gorithm acquires skills that represent shared structures across changing maze na
vigation and locomotion environments. Furthermore, we demonstrate that our skill
s are more useful than baselines on downstream tasks.
****

Generalization Analysis for Contrastive Representation Learning
Yunwen Lei, Tianbao Yang, Yiming Ying, Ding-Xuan Zhou
Recently, contrastive learning has found impressive success in advancing the sta
te of the art in solving various machine learning tasks. However, the existing g
eneralization analysis is very limited or even not meaningful. In particular, th
e existing generalization error bounds depend linearly on the number k of nega
tive examples while it was widely shown in practice that choosing a large k is
 necessary to guarantee good generalization of contrastive learning in downstrea
m tasks. In this paper, we establish novel generalization bounds for contrastive
 learning which do not depend on k, up to logarithmic terms. Our analysis uses
 structural results on empirical covering numbers and Rademacher complexities to
 exploit the Lipschitz continuity of loss functions. For self-bounding Lipschitz
 loss functions, we further improve our results by developing optimistic bounds
which imply fast rates in a low noise condition. We apply our results to learnin
g with both linear representation and nonlinear representation by deep neural ne
tworks, for both of which we derive Rademacher complexity bounds to get improved
 generalization bounds.
****

Learning Control by Iterative Inversion
Gal Leibovich, Guy Jacob, Or Avner, Gal Novik, Aviv Tamar
We propose iterative inversion - an algorithm for learning an inverse function w

ithout input-output pairs, but only with samples from the desired output distrib ution and access to the forward function. The key challenge is a distribution sh ift between the desired outputs and the outputs of an initial random guess, and we prove that iterative inversion can steer the learning correctly, under rather strict conditions on the function. We apply iterative inversion to learn contro l. Our input is a set of demonstrations of desired behavior, given as video embe ddings of trajectories (without actions), and our method iteratively learns to i mitate trajectories generated by the current policy, perturbed by random explora tion noise. Our approach does not require rewards, and only employs supervised l earning, which can be easily scaled to use state-of-the-art trajectory embedding techniques and policy representations. Indeed, with a VQ-VAE embedding, and a t ransformer-based policy, we demonstrate non-trivial continuous control on severa l tasks (videos available at https://sites.google.com/view/iter-inver). Further, we report an improved performance on imitating diverse behaviors compared to re ward based methods.

****

Sampling-Based Accuracy Testing of Posterior Estimators for General Inference

Pablo Lemos, Adam Coogan, Yashar Hezaveh, Laurence Perreault-Levasseur

Parameter inference, i.e. inferring the posterior distribution of the parameters of a statistical model given some data, is a central problem to many scientific disciplines. Posterior inference with generative models is an alternative to me thods such as Markov Chain Monte Carlo, both for likelihood-based and simulation -based inference. However, assessing the accuracy of posteriors encoded in gener ative models is not straightforward. In this paper, we introduce "Tests of Accur acy with Random Points" (TARP) coverage testing as a method to estimate coverage probabilities of generative posterior estimators. Our method differs from previ ously-existing coverage-based methods, which require posterior evaluations. We p rove that our approach is necessary and sufficient to show that a posterior esti mator is accurate. We demonstrate the method on a variety of synthetic examples, and show that TARP can be used to test the results of posterior inference analy ses in high-dimensional spaces. We also show that our method can detect inaccura te inferences in cases where existing methods fail.

****

Fast Inference from Transformers via Speculative Decoding

Yaniv Leviathan, Matan Kalman, Yossi Matias

Inference from large autoregressive models like Transformers is slow - decoding K tokens takes K serial runs of the model. In this work we introduce speculative decoding - an algorithm to sample from autoregressive models faster without any changes to the outputs, by computing several tokens in parallel. At the heart o f our approach lie the observations that (1) hard language-modeling tasks often include easier subtasks that can be approximated well by more efficient models, and (2) using speculative execution and a novel sampling method, we can make exa ct decoding from the large models faster, by running them in parallel on the out puts of the approximation models, potentially generating several tokens concurre ntly, and without changing the distribution. Our method can accelerate existing off-the-shelf models without retraining or architecture changes. We demonstrate it on T5-XXL and show a 2X-3X acceleration compared to the standard T5X implemen tation, with identical outputs.

****

Efficient Rate Optimal Regret for Adversarial Contextual MDPs Using Online Funct ion Approximation

Orin Levy, Alon Cohen, Asaf Cassel, Yishay Mansour

We present the OMG-CMDP! algorithm for regret minimization in adversarial Contex tual MDPs. The algorithm operates under the minimal assumptions of realizable fu nction class and access to online least squares and log loss regression oracles. Our algorithm is efficient (assuming efficient online regression oracles), simp le and robust to approximation errors. It enjoys an $\widetilde{O}(H^{2.5} \sqrt { T|S||A| (\mathcal{R}_{TH}(\mathcal{O}) + H \log(\delta^{-1}))})$ regret guar antee, with T being the number of episodes, S the state space, A the actio n space, H the horizon and $\mathcal{R}_{TH}(\mathcal{O}) = \mathcal{R}_{TH}(\

mathcal{O}_{sq}^\mathcal{F}) + \mathcal{R}_{TH}(\mathcal{O}_{log}^\mathcal{P})$ is the sum of the square and log-loss regression oracles' regret, used to approximate the context-dependent rewards and dynamics, respectively. To the best of our knowledge, our algorithm is the first efficient rate optimal regret minimization algorithm for adversarial CMDPs that operates under the minimal standard assumption of online function approximation.

****

GLOBE-CE: A Translation Based Approach for Global Counterfactual Explanations

Dan Ley, Saumitra Mishra, Daniele Magazzeni

Counterfactual explanations have been widely studied in explainability, with a range of application dependent methods prominent in fairness, recourse and model understanding. The major shortcoming associated with these methods, however, is their inability to provide explanations beyond the local or instance-level. While many works touch upon the notion of a global explanation, typically suggesting to aggregate masses of local explanations in the hope of ascertaining global properties, few provide frameworks that are both reliable and computationally tractable. Meanwhile, practitioners are requesting more efficient and interactive explainability tools. We take this opportunity to propose Global & Efficient Counterfactual Explanations (GLOBE-CE), a flexible framework that tackles the reliability and scalability issues associated with current state-of-the-art, particularly on higher dimensional datasets and in the presence of continuous features. Furthermore, we provide a unique mathematical analysis of categorical feature translations, utilising it in our method. Experimental evaluation with publicly available datasets and user studies demonstrate that GLOBE-CE performs significantly better than the current state-of-the-art across multiple metrics (e.g., speed, reliability).

****

TIPS: Topologically Important Path Sampling for Anytime Neural Networks

Guihong Li, Kartikeya Bhardwaj, Yuedong Yang, Radu Marculescu

Anytime neural networks (AnytimeNNs) are a promising solution to adaptively adjust the model complexity at runtime under various hardware resource constraints. However, the manually-designed AnytimeNNs are biased by designers' prior experience and thus provide sub-optimal solutions. To address the limitations of existing hand-crafted approaches, we first model the training process of AnytimeNNs as a discrete-time Markov chain (DTMC) and use it to identify the paths that contribute the most to the training of AnytimeNNs. Based on this new DTMC-based analysis, we further propose TIPS, a framework to automatically design AnytimeNNs under various hardware constraints. Our experimental results show that TIPS can improve the convergence rate and test accuracy of AnytimeNNs. Compared to the existing AnytimeNNs approaches, TIPS improves the accuracy by 2%-6.6% on multiple datasets and achieves SOTA accuracy-FLOPs tradeoffs.

****

MAHALO: Unifying Offline Reinforcement Learning and Imitation Learning from Observations

Anqi Li, Byron Boots, Ching-An Cheng

We study a new paradigm for sequential decision making, called offline policy learning from observations (PLfO). Offline PLfO aims to learn policies using datasets with substandard qualities: 1) only a subset of trajectories is labeled with rewards, 2) labeled trajectories may not contain actions, 3) labeled trajectories may not be of high quality, and 4) the data may not have full coverage. Such imperfection is common in real-world learning scenarios, and offline PLfO encompasses many existing offline learning setups, including offline imitation learning (IL), offline IL from observations (ILfO), and offline reinforcement learning (RL). In this work, we present a generic approach to offline PLfO, called Modality-agnostic Adversarial Hypothesis Adaptation for Learning from Observations (MAHALO). Built upon the pessimism concept in offline RL, MAHALO optimizes the policy using a performance lower bound that accounts for uncertainty due to the dataset's insufficient coverage. We implement this idea by adversarially training data-consistent critic and reward functions, which forces the learned policy to be robust to data deficiency. We show that MAHALO consistently outperforms or matc

hes specialized algorithms across a variety of offline PLfO tasks in theory and experiments. Our code is available at https://github.com/AnqiLi/mahalo.

****

Internet Explorer: Targeted Representation Learning on the Open Web
Alexander Cong Li, Ellis Langham Brown, Alexei A Efros, Deepak Pathak
Vision models typically rely on fine-tuning general-purpose models pre-trained o n large, static datasets. These general-purpose models only capture the knowledg e within their pre-training datasets, which are tiny, out-of-date snapshots of t he Internet—where billions of images are uploaded each day. We suggest an altern ate approach: rather than hoping our static datasets transfer to our desired tas ks after large-scale pre-training, we propose dynamically utilizing the Internet to quickly train a small-scale model that does extremely well on a target datas et. Our approach, called Internet Explorer, explores the web in a self-supervise d manner to progressively find relevant examples that improve performance on a d esired target dataset. It cycles between searching for images on the Internet wi th text queries, self-supervised training on downloaded images, determining whic h images were useful, and prioritizing what to search for next. We evaluate Inte rnet Explorer across several datasets and show that it outperforms or matches CL IP oracle performance using just a single GPU desktop to actively query the Inte rnet for 30-40 hours.

****

Prototype-oriented unsupervised anomaly detection for multivariate time series
Yuxin Li, Wenchao Chen, Bo Chen, Dongsheng Wang, Long Tian, Mingyuan Zhou
Unsupervised anomaly detection (UAD) of multivariate time series (MTS) aims to l earn robust representations of normal multivariate temporal patterns. Existing U AD methods try to learn a fixed set of mappings for each MTS, entailing expensiv e computation and limited model adaptation. To address this pivotal issue, we pr opose a prototype-oriented UAD (PUAD) method under a probabilistic framework. Sp ecifically, instead of learning the mappings for each MTS, the proposed PUAD vie ws multiple MTSs as the distribution over a group of prototypes, which are extra cted to represent a diverse set of normal patterns. To learn and regulate the pr ototypes, PUAD introduces a reconstruction-based unsupervised anomaly detection approach, which incorporates a prototype-oriented optimal transport method into a Transformer-powered probabilistic dynamical generative framework. Leveraging m eta-learned transferable prototypes, PUAD can achieve high model adaptation capa city for new MTSs. Experiments on five public MTS datasets all verify the effect iveness of the proposed UAD method.

****

Learning Preconditioners for Conjugate Gradient PDE Solvers
Yichen Li, Peter Yichen Chen, Tao Du, Wojciech Matusik
Efficient numerical solvers for partial differential equations empower science a nd engineering. One commonly employed numerical solver is the preconditioned con jugate gradient (PCG) algorithm, whose performance is largely affected by the pr econditioner quality. However, designing high-performing preconditioner with tra ditional numerical methods is highly non-trivial, often requiring problem-specif ic knowledge and meticulous matrix operations. We present a new method that leve rages learning-based approach to obtain an approximate matrix factorization to t he system matrix to be used as a preconditioner in the context of PCG solvers. O ur high-level intuition comes from the shared property between preconditioners a nd network-based PDE solvers that excels at obtaining approximate solutions at a low computational cost. Such observation motivates us to represent precondition ers as graph neural networks (GNNs). In addition, we propose a new loss function that rewrites traditional preconditioner metrics to incorporate inductive bias from PDE data distributions, enabling effective training of high-performing prec onditioners. We conduct extensive experiments to demonstrate the efficacy and ge neralizability of our proposed approach on solving various 2D and 3D linear seco nd-order PDEs.

****

Parallel Q-Learning: Scaling Off-policy Reinforcement Learning under Massively Parallel Simulation

Zechu Li, Tao Chen, Zhang-Wei Hong, Anurag Ajay, Pulkit Agrawal
Reinforcement learning is time-consuming for complex tasks due to the need for large amounts of training data. Recent advances in GPU-based simulation, such as Isaac Gym, have sped up data collection thousands of times on a commodity GPU. Most prior works have used on-policy methods like PPO due to their simplicity and easy-to-scale nature. Off-policy methods are more sample-efficient, but challenging to scale, resulting in a longer wall-clock training time. This paper presents a novel Parallel Q-Learning (PQL) scheme that outperforms PPO in terms of wall-clock time and maintains superior sample efficiency. The driving force lies in the parallelization of data collection, policy function learning, and value function learning. Different from prior works on distributed off-policy learning, such as Apex, our scheme is designed specifically for massively parallel GPU-based simulation and optimized to work on a single workstation. In experiments, we demonstrate the capability of scaling up Q-learning methods to tens of thousands of parallel environments and investigate important factors that can affect learning speed, including the number of parallel environments, exploration strategies, batch size, GPU models, etc. The code is available at https://github.com/Improbable-AI/pql.
****

Minimum Width of Leaky-ReLU Neural Networks for Uniform Universal Approximation
Li'Ang Li, Yifei Duan, Guanghua Ji, Yongqiang Cai
The study of universal approximation properties (UAP) for neural networks (NN) has a long history. When the network width is unlimited, only a single hidden layer is sufficient for UAP. In contrast, when the depth is unlimited, the width for UAP needs to be not less than the critical width $w^*_{\min}=\max(d_x,d_y)$, where d_x and d_y are the dimensions of the input and output, respectively. Recently, (Cai, 2022) shows that a leaky-ReLU NN with this critical width can achieve UAP for L^p functions on a compact domain $\mathcal{K}$, i.e., the UAP for $L^p(\mathcal{K},\mathbb{R}^{d_y})$. This paper examines a uniform UAP for the function class $C(\mathcal{K},\mathbb{R}^{d_y})$ and gives the exact minimum width of the leaky-ReLU NN as $w_{\min}=\max(d_x+1,d_y)+1_{d_y=d_x+1}$, which involves the effects of the output dimensions. To obtain this result, we propose a novel lift-flow-discretization approach that shows that the uniform UAP has a deep connection with topological theory.
****

FAIRER: Fairness as Decision Rationale Alignment
Tianlin Li, Qing Guo, Aishan Liu, Mengnan Du, Zhiming Li, Yang Liu
Deep neural networks (DNNs) have made significant progress, but often suffer from fairness issues, as deep models typically show distinct accuracy differences among certain subgroups (e.g., males and females). Existing research addresses this critical issue by employing fairness-aware loss functions to constrain the last-layer outputs and directly regularize DNNs. Although the fairness of DNNs is improved, it is unclear how the trained network makes a fair prediction, which limits future fairness improvements. In this paper, we investigate fairness from the perspective of decision rationale and define the parameter parity score to characterize the fair decision process of networks by analyzing neuron influence in various subgroups. Extensive empirical studies show that the unfair issue could arise from the unaligned decision rationales of subgroups. Existing fairness regularization terms fail to achieve decision rationale alignment because they only constrain last-layer outputs while ignoring intermediate neuron alignment. To address the issue, we formulate the fairness as a new task, i.e., decision rationale alignment that requires DNNs' neurons to have consistent responses on subgroups at both intermediate processes and the final prediction. To make this idea practical during optimization, we relax the naive objective function and propose gradient-guided parity alignment, which encourages gradient-weighted consistency of neurons across subgroups. Extensive experiments on a variety of datasets show that our method can significantly enhance fairness while sustaining a high level of accuracy and outperforming other approaches by a wide margin.
****

RACE: Improve Multi-Agent Reinforcement Learning with Representation Asymmetry a

nd Collaborative Evolution
Pengyi Li, Jianye Hao, Hongyao Tang, Yan Zheng, Xian Fu
Multi-Agent Reinforcement Learning (MARL) has demonstrated its effectiveness in learning collaboration, but it often struggles with low-quality reward signals and high non-stationarity. In contrast, Evolutionary Algorithm (EA) has shown better convergence, robustness, and signal quality insensitivity. This paper introduces a hybrid framework, Representation Asymmetry and Collaboration Evolution (RACE), which combines EA and MARL for efficient collaboration. RACE maintains a MARL team and a population of EA teams. To enable efficient knowledge sharing and policy exploration, RACE decomposes the policies of different teams controlling the same agent into a shared nonlinear observation representation encoder and individual linear policy representations. To address the partial observation issue, we introduce Value-Aware Mutual Information Maximization to enhance the shared representation with useful information about superior global states. EA evolves the population using novel agent-level crossover and mutation operators, offering diverse experiences for MARL. Concurrently, MARL optimizes its policies and injects them into the population for evolution. The experiments on challenging continuous and discrete tasks demonstrate that RACE significantly improves the basic algorithms, consistently outperforming other algorithms. Our code is available at https://github.com/yeshenpy/RACE.
****

Adversarial Collaborative Learning on Non-IID Features
Qinbin Li, Bingsheng He, Dawn Song
Federated Learning (FL) has been a popular approach to enable collaborative learning on multiple parties without exchanging raw data. However, the model performance of FL may degrade a lot due to non-IID data. While many FL algorithms focus on non-IID labels, FL on non-IID features has largely been overlooked. Different from typical FL approaches, the paper proposes a new learning concept called ADCOL (Adversarial Collaborative Learning) for non-IID features. Instead of adopting the widely used model-averaging scheme, ADCOL conducts training in an adversarial way: the server aims to train a discriminator to distinguish the representations of the parties, while the parties aim to generate a common representation distribution. Our experiments show that ADCOL achieves better performance than state-of-the-art FL algorithms on non-IID features.
****

Near-optimal Conservative Exploration in Reinforcement Learning under Episode-wise Constraints
Donghao Li, Ruiquan Huang, Cong Shen, Jing Yang
This paper investigates conservative exploration in reinforcement learning where the performance of the learning agent is guaranteed to be above a certain threshold throughout the learning process. It focuses on the tabular episodic Markov Decision Process (MDP) setting that has finite states and actions. With the knowledge of an existing safe baseline policy, an algorithms termed as StepMix is proposed to balance the exploitation and exploration while ensuring that the conservative constraint is never violated in each episode with high probability. StepMix features a unique design of a mixture policy that adaptively and smoothly interpolates between the baseline policy and the optimistic policy. Theoretical analysis shows that StepMix achieves near-optimal regret order as in the constraint-free setting, indicating that obeying the stringent episode-wise conservative constraint does not compromise the learning performance. Besides, a randomization based EpsMix algorithm is also proposed and shown the achieve the same performance as StepMix. The algorithm design and theoretical analysis are further extended to the setting where the baseline policy is not given a priori but must be learned from an offline dataset, and it is proved that similar conservative guarantee and regret can be achieved if the offline dataset is sufficiently large. Experiment results corroborate the theoretical analysis and demonstrate the effectiveness of the proposed conservative exploration strategies.
****

Transformers as Algorithms: Generalization and Stability in In-context Learning
Yingcong Li, Muhammed Emrullah Ildiz, Dimitris Papailiopoulos, Samet Oymak

In-context learning (ICL) is a type of prompting where a transformer model operates on a sequence of (input, output) examples and performs inference on-the-fly. In this work, we formalize in-context learning as an algorithm learning problem where a transformer model implicitly constructs a hypothesis function at inference-time. We first explore the statistical aspects of this abstraction through the lens of multitask learning: We obtain generalization bounds for ICL when the input prompt is (1) a sequence of i.i.d. (input, label) pairs or (2) a trajectory arising from a dynamical system. The crux of our analysis is relating the excess risk to the stability of the algorithm implemented by the transformer. We characterize when transformer/attention architecture provably obeys the stability condition and also provide empirical verification. For generalization on unseen tasks, we identify an inductive bias phenomenon in which the transfer learning risk is governed by the task complexity and the number of MTL tasks in a highly predictable manner. Finally, we provide numerical evaluations that (1) demonstrate transformers can indeed implement near-optimal algorithms on classical regression problems with i.i.d. and dynamic data, (2) provide insights on stability, and (3) verify our theoretical predictions.

****

Improving Hyperparameter Learning under Approximate Inference in Gaussian Process Models

Rui Li, S. T. John, Arno Solin

Approximate inference in Gaussian process (GP) models with non-conjugate likelihoods gets entangled with the learning of the model hyperparameters. We improve hyperparameter learning in GP models and focus on the interplay between variational inference (VI) and the learning target. While VI’s lower bound to the marginal likelihood is a suitable objective for inferring the approximate posterior, we show that a direct approximation of the marginal likelihood as in Expectation Propagation (EP) is a better learning objective for hyperparameter optimization. We design a hybrid training procedure to bring the best of both worlds: it leverages conjugate-computation VI for inference and uses an EP-like marginal likelihood approximation for hyperparameter learning. We compare VI, EP, Laplace approximation, and our proposed training procedure and empirically demonstrate the effectiveness of our proposal across a wide range of data sets.

****

Local Vertex Colouring Graph Neural Networks

Shouheng Li, Dongwoo Kim, Qing Wang

In recent years, there has been a significant amount of research focused on expanding the expressivity of Graph Neural Networks (GNNs) beyond the Weisfeiler-Leman (1-WL) framework. While many of these studies have yielded advancements in expressivity, they have frequently come at the expense of decreased efficiency or have been restricted to specific types of graphs. In this study, we investigate the expressivity of GNNs from the perspective of graph search. Specifically, we propose a new vertex colouring scheme and demonstrate that classical search algorithms can efficiently compute graph representations that extend beyond the 1-WL. We show the colouring scheme inherits useful properties from graph search that can help solve problems like graph biconnectivity. Furthermore, we show that under certain conditions, the expressivity of GNNs increases hierarchically with the radius of the search neighbourhood. To further investigate the proposed scheme, we develop a new type of GNN based on two search strategies, breadth-first search and depth-first search, highlighting the graph properties they can capture on top of 1-WL. Our code is available at https://github.com/seanli3/lvc.

****

Analysis of Error Feedback in Federated Non-Convex Optimization with Biased Compression: Fast Convergence and Partial Participation

Xiaoyun Li, Ping Li

In practical federated learning (FL) systems, the communication cost between the clients and the central server can often be a bottleneck. In this paper, we focus on biased gradient compression in non-convex FL problems. In the classical distributed learning, the method of error feedback (EF) is a common technique to remedy the downsides of biased gradient compression, but the performance of EF in

FL still lacks systematic investigation. In this work, we study a compressed FL scheme with error feedback, named Fed-EF, with two variants depending on the global model optimizer. While directly applying biased compression in FL leads to poor convergence, we show that Fed-EF is able to match the convergence rate of the full-precision FL counterpart with a linear speedup w.r.t. the number of clients. Experiments verify that Fed-EF achieves the same performance as the full-precision FL approach, at the substantially reduced communication cost. Moreover, we develop a new analysis of the EF under partial participation (PP), an important scenario in FL. Under PP, the convergence rate of Fed-EF exhibits an extra slow-down factor due to a so-called “stale error compensation” effect, which is also justified in our experiments. Our results provide insights on a theoretical limitation of EF, and possible directions for improvements.

****

How Do Transformers Learn Topic Structure: Towards a Mechanistic Understanding

Yuchen Li, Yuanzhi Li, Andrej Risteski

While the successes of transformers across many domains are indisputable, accurate understanding of the learning mechanics is still largely lacking. Their capabilities have been probed on benchmarks which include a variety of structured and reasoning tasks—but mathematical understanding is lagging substantially behind. Recent lines of work have begun studying representational aspects of this question: that is, the size/depth/complexity of attention-based networks to perform certain tasks. However, there is no guarantee the learning dynamics will converge to the constructions proposed. In our paper, we provide fine-grained mechanistic understanding of how transformers learn “semantic structure”, understood as capturing co-occurrence structure of words. Precisely, we show, through a combination of mathematical analysis and experiments on Wikipedia data and synthetic data modeled by Latent Dirichlet Allocation (LDA), that the embedding layer and the self-attention layer encode the topical structure. In the former case, this manifests as higher average inner product of embeddings between same-topic words. In the latter, it manifests as higher average pairwise attention between same-topic words. The mathematical results involve several assumptions to make the analysis tractable, which we verify on data, and might be of independent interest as well.

****

BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models

Junnan Li, Dongxu Li, Silvio Savarese, Steven Hoi

The cost of vision-and-language pre-training has become increasingly prohibitive due to end-to-end training of large-scale models. This paper proposes BLIP-2, a generic and efficient pre-training strategy that bootstraps vision-language pre-training from off-the-shelf frozen pre-trained image encoders and frozen large language models. BLIP-2 bridges the modality gap with a lightweight Querying Transformer, which is pre-trained in two stages. The first stage bootstraps vision-language representation learning from a frozen image encoder. The second stage bootstraps vision-to-language generative learning from a frozen language model. BLIP-2 achieves state-of-the-art performance on various vision-language tasks, despite having significantly fewer trainable parameters than existing methods. For example, our model outperforms Flamingo80B by 8.7% on zero-shot VQAv2 with 54x fewer trainable parameters. We also demonstrate the model’s emerging capabilities of zero-shot image-to-text generation that can follow natural language instructions.

****

Nearly Optimal Algorithms with Sublinear Computational Complexity for Online Kernel Regression

Junfan Li, Shizhong Liao

The trade-off between regret and computational cost is a fundamental problem for online kernel regression, and previous algorithms worked on the trade-off can not keep optimal regret bounds at a sublinear computational complexity. In this paper, we propose two new algorithms, AOGD-ALD and NONS-ALD, which can keep nearly optimal regret bounds at a sublinear computational complexity, and give suffic

ient conditions under which our algorithms work. Both algorithms dynamically mai ntain a group of nearly orthogonal basis used to approximate the kernel mapping, and keep nearly optimal regret bounds by controlling the approximate error. The number of basis depends on the approximate error and the decay rate of eigenval ues of the kernel matrix. If the eigenvalues decay exponentially, then AOGD-ALD and NONS-ALD separately achieves a regret of $O(\sqrt{L(f)})$ and $O(\mathrm{d}_ {\mathrm{eff}}(\mu)\ln{T})$ at a computational complexity in $O(\ln^2{T})$. If t he eigenvalues decay polynomially with degree $p\geq 1$, then our algorithms kee p the same regret bounds at a computational complexity in $o(T)$ in the case of $p>4$ and $p\geq 10$, respectively. $L(f)$ is the cumulative losses of f and $ \mathrm{d}_{\mathrm{eff}}(\mu)$ is the effective dimension of the problem. The t wo regret bounds are nearly optimal and are not comparable.

****

Revisiting Weighted Aggregation in Federated Learning with Neural Networks

Zexi Li, Tao Lin, Xinyi Shang, Chao Wu

In federated learning (FL), weighted aggregation of local models is conducted to generate a global model, and the aggregation weights are normalized (the sum of weights is 1) and proportional to the local data sizes. In this paper, we revis it the weighted aggregation process and gain new insights into the training dyna mics of FL. First, we find that the sum of weights can be smaller than 1, causin g global weight shrinking effect (analogous to weight decay) and improving gener alization. We explore how the optimal shrinking factor is affected by clients' d ata heterogeneity and local epochs. Second, we dive into the relative aggregatio n weights among clients to depict the clients' importance. We develop client coh erence to study the learning dynamics and find a critical point that exists. Bef ore entering the critical point, more coherent clients play more essential roles in generalization. Based on the above insights, we propose an effective method for Federated Learning with Learnable Aggregation Weights, named as FedLAW. Exte nsive experiments verify that our method can improve the generalization of the g lobal model by a large margin on different datasets and models.

****

Distribution-dependent McDiarmid-type Inequalities for Functions of Unbounded In teraction

Shaojie Li, Yong Liu

The concentration of measure inequalities serves an essential role in statistics and machine learning. This paper gives unbounded analogues of the McDiarmid-typ e exponential inequalities for three popular classes of distributions, namely su b-Gaussian, sub-exponential and heavy-tailed distributions. The inequalities in the sub-Gaussian and sub-exponential cases are distribution-dependent compared w ith the recent results, and the inequalities in the heavy-tailed case are not av ailable in the previous works. The usefulness of the inequalities is illustrated through applications to the sample mean, U-statistics and V-statistics.

****

Optimal Convergence Rates for Agnostic Nyström Kernel Learning

Jian Li, Yong Liu, Weiping Wang

Nyström low-rank approximation has shown great potential in processing large-sca le kernel matrix and neural networks. However, there lacks a unified analysis fo r Nyström approximation, and the asymptotical minimax optimality for Nyström met hods usually require a strict condition, assuming that the target regression lie s exactly in the hypothesis space. In this paper, to tackle these problems, we p rovide a refined generalization analysis for Nyström approximation in the agnost ic setting, where the target regression may be out of the hypothesis space. Spec ifically, we show Nyström approximation can still achieve the capacity-dependent optimal rates in the agnostic setting. To this end, we first prove the capacity -dependent optimal guarantees of Nyström approximation with the standard uniform sampling, which covers both loss functions and applies to some agnostic setting s. Then, using data-dependent sampling, for example, leverage scores sampling, w e derive the capacity-dependent optimal rates that apply to the whole range of t he agnostic setting. To our best knowledge, the capacity-dependent optimality fo r the whole range of the agnostic setting is first achieved and novel in Nyström

approximation.
****
Reconstructive Neuron Pruning for Backdoor Defense
Yige Li, Xixiang Lyu, Xingjun Ma, Nodens Koren, Lingjuan Lyu, Bo Li, Yu-Gang Jia
ng
Deep neural networks (DNNs) have been found to be vulnerable to backdoor attacks
, raising security concerns about their deployment in mission-critical applicati
ons. While existing defense methods have demonstrated promising results, it is s
till not clear how to effectively remove backdoor-associated neurons in backdoor
ed DNNs. In this paper, we propose a novel defense called Reconstructive Neuron
Pruning (RNP) to expose and prune backdoor neurons via an unlearning and then re
covering process. Specifically, RNP first unlearns the neurons by maximizing the
model's error on a small subset of clean samples and then recovers the neurons
by minimizing the model's error on the same data. In RNP, unlearning is operated
at the neuron level while recovering is operated at the filter level, forming a
n asymmetric reconstructive learning procedure. We show that such an asymmetric
process on only a few clean samples can effectively expose and prune the backdoo
r neurons implanted by a wide range of attacks, achieving a new state-of-the-art
defense performance. Moreover, the unlearned model at the intermediate step of
our RNP can be directly used to improve other backdoor defense tasks including b
ackdoor removal, trigger recovery, backdoor label detection, and backdoor sample
detection. Code is available at https://github.com/bboylyg/RNP.
****
Meta Learning of Interface Conditions for Multi-Domain Physics-Informed Neural N
etworks
Shibo Li, Michael Penwarden, Yiming Xu, Conor Tillinghast, Akil Narayan, Mike Ki
rby, Shandian Zhe
Physics-informed neural networks (PINNs) are emerging as popular mesh-free solve
rs for partial differential equations (PDEs). Recent extensions decompose the do
main, apply different PINNs to solve the problem in each subdomain, and stitch t
he subdomains at the interface. Thereby, they can further alleviate the problem
complexity, reduce the computational cost, and allow parallelization. However, t
he performance of multi-domain PINNs is sensitive to the choice of the interface
conditions. While quite a few conditions have been proposed, there is no sugges
tion about how to select the conditions according to specific problems. To addre
ss this gap, we propose META Learning of Interface Conditions (METALIC), a simpl
e, efficient yet powerful approach to dynamically determine appropriate interfac
e conditions for solving a family of parametric PDEs. Specifically, we develop t
wo contextual multi-arm bandit (MAB) models. The first one applies to the entire
training course, and online updates a Gaussian process (GP) reward that given t
he PDE parameters and interface conditions predicts the performance. We prove a
sub-linear regret bound for both UCB and Thompson sampling, which in theory guar
antees the effectiveness of our MAB. The second one partitions the training into
two stages, one is the stochastic phase and the other deterministic phase; we u
pdate a GP reward for each phase to enable different condition selections at the
two stages to further bolster the flexibility and performance. We have shown th
e advantage of METALIC on four bench-mark PDE families.
****
Deep Anomaly Detection under Labeling Budget Constraints
Aodong Li, Chen Qiu, Marius Kloft, Padhraic Smyth, Stephan Mandt, Maja Rudolph
Selecting informative data points for expert feedback can significantly improve
the performance of anomaly detection (AD) in various contexts, such as medical d
iagnostics or fraud detection. In this paper, we determine a set of theoretical
conditions under which anomaly scores generalize from labeled queries to unlabel
ed data. Motivated by these results, we propose a data labeling strategy with op
timal data coverage under labeling budget constraints. In addition, we propose a
new learning framework for semi-supervised AD. Extensive experiments on image,
tabular, and video data sets show that our approach results in state-of-the-art
semi-supervised AD performance under labeling budget constraints.
****

On the Initialization of Graph Neural Networks
Jiahang Li, Yakun Song, Xiang Song, David Wipf
Graph Neural Networks (GNNs) have displayed considerable promise in graph representation learning across various applications. The core learning process requires the initialization of model weight matrices within each GNN layer, which is typically accomplished via classic initialization methods such as Xavier initialization. However, these methods were originally motivated to stabilize the variance of hidden embeddings and gradients across layers of Feedforward Neural Networks (FNNs) and Convolutional Neural Networks (CNNs) to avoid vanishing gradients and maintain steady information flow. In contrast, within the GNN context classical initializations disregard the impact of the input graph structure and message passing on variance. In this paper, we analyze the variance of forward and backward propagation across GNN layers and show that the variance instability of GNN initializations comes from the combined effect of the activation function, hidden dimension, graph structure and message passing. To better account for these influence factors, we propose a new initialization method for Variance Instability Reduction within GNN Optimization (Virgo), which naturally tends to equate forward and backward variances across successive layers. We conduct comprehensive experiments on 15 datasets to show that Virgo can lead to superior model performance and more stable variance at initialization on node classification, link prediction and graph classification tasks.
****
Federated Adversarial Learning: A Framework with Convergence Analysis
Xiaoxiao Li, Zhao Song, Jiaming Yang
Federated learning (FL) is a trending training paradigm to utilize decentralized training data. FL allows clients to update model parameters locally for several epochs, then share them to a global model for aggregation. This training paradigm with multi-local step updating before aggregation exposes unique vulnerabilities to adversarial attacks. Adversarial training is a popular and effective method to improve the robustness of networks against adversaries. In this work, we formulate a general form of federated adversarial learning (FAL) that is adapted from adversarial learning in the centralized setting. On the client side of FL training, FAL has an inner loop to generate adversarial samples for adversarial training and an outer loop to update local model parameters. On the server side, FAL aggregates local model updates and broadcast the aggregated model. We design a global robust training loss and formulate FAL training as a min-max optimization problem. Unlike the convergence analysis in classical centralized training that relies on the gradient direction, it is significantly harder to analyze the convergence in FAL for three reasons: 1) the complexity of min-max optimization, 2) model not updating in the gradient direction due to the multi-local updates on the client-side before aggregation and 3) inter-client heterogeneity. We address these challenges by using appropriate gradient approximation and coupling techniques and present the convergence analysis in the over-parameterized regime. Our main result theoretically shows that the minimum loss under our algorithm can converge to ϵ small with chosen learning rate and communication rounds. It is noteworthy that our analysis is feasible for non-IID clients.
****
How Powerful are Shallow Neural Networks with Bandlimited Random Weights?
Ming Li, Sho Sonoda, Feilong Cao, Yu Guang Wang, Jiye Liang
We investigate the expressive power of depth-2 bandlimited random neural networks. A random net is a neural network where the hidden layer parameters are frozen with random assignment, and only the output layer parameters are trained by loss minimization. Using random weights for a hidden layer is an effective method to avoid non-convex optimization in standard gradient descent learning. It has also been adopted in recent deep learning theories. Despite the well-known fact that a neural network is a universal approximator, in this study, we mathematically show that when hidden parameters are distributed in a bounded domain, the network may not achieve zero approximation error. In particular, we derive a new nontrivial approximation error lower bound. The proof utilizes the technique of ridgelet analysis, a harmonic analysis method designed for neural networks. This me

thod is inspired by fundamental principles in classical signal processing, specifically the idea that signals with limited bandwidth may not always be able to perfectly reconstruct the original signal. We corroborate our theoretical results with various simulation studies, and generally, two main take-home messages are offered: (i) Not any distribution for selecting random weights is feasible to build a universal approximator; (ii) A suitable assignment of random weights exists but to some degree is associated with the complexity of the target function.
****
Efficient Quantum Algorithms for Quantum Optimal Control
Xiantao Li, Chunhao Wang
In this paper, we present efficient quantum algorithms that are exponentially faster than classical algorithms for solving the quantum optimal control problem. This problem involves finding the control variable that maximizes a physical quantity at time T, where the system is governed by a time-dependent Schrödinger equation. This type of control problem also has an intricate relation with machine learning. Our algorithms are based on a time-dependent Hamiltonian simulation method and a fast gradient-estimation algorithm. We also provide a comprehensive error analysis to quantify the total error from various steps, such as the finite-dimensional representation of the control function, the discretization of the Schrödinger equation, the numerical quadrature, and optimization. Our quantum algorithms require fault-tolerant quantum computers.
****
Low-Switching Policy Gradient with Exploration via Online Sensitivity Sampling
Yunfan Li, Yiran Wang, Yu Cheng, Lin Yang
Policy optimization methods are powerful algorithms in Reinforcement Learning (RL) for their flexibility to deal with policy parameterization and ability to handle model misspecification. However, these methods usually suffer from slow convergence rates and poor sample complexity. Hence it is important to design provably sample efficient algorithms for policy optimization. Yet, recent advances for this problems have only been successful in tabular and linear setting, whose benign structures cannot be generalized to non-linearly parameterized policies. In this paper, we address this problem by leveraging recent advances in value-based algorithms, including bounded eluder-dimension and online sensitivity sampling, to design a low-switching sample-efficient policy optimization algorithm, LPO, with general non-linear function approximation. We show that, our algorithm obtains an ε-optimal policy with only $\widetilde{O}(\frac{\text{poly}(d)}{\varepsilon^3})$ samples, where ε is the suboptimality gap and d is a complexity measure of the function class approximating the policy. This drastically improves previously best-known sample bound for policy optimization algorithms, $\widetilde{O}(\frac{\text{poly}(d)}{\varepsilon^8})$. Moreover, we empirically test our theory with deep neural nets to show the benefits of the theoretical inspiration.
****
Hierarchical Diffusion for Offline Decision Making
Wenhao Li, Xiangfeng Wang, Bo Jin, Hongyuan Zha
Offline reinforcement learning typically introduces a hierarchical structure to solve the long-horizon problem so as to address its thorny issue of variance accumulation. Problems of deadly triad, limited data and reward sparsity, however, still remain, rendering the design of effective, hierarchical offline RL algorithms for general-purpose policy learning a formidable challenge. In this paper, we first formulate the problem of offline long-horizon decision-$\mathbf{M}$ak$\mathbf{I}$ng from the perspective of conditional generative modeling by incorporating goals into the control-as-inference graphic models. A $\mathbf{H}$ierarchical trajectory-level $\mathbf{D}$iffusion probabilistic model is then proposed with classifier-free guidance. HDMI employs a cascade framework that utilizes the reward-conditional goal diffuser for the subgoal discovery and the goal-conditional trajectory diffuser for generating the corresponding action sequence of subgoals. Planning-based subgoal extraction and transformer-based diffusion are employed to deal with the sub-optimal data pollution and long-range subgoal dependencies in the goal diffusion. Numerical experiments verify the advantages of HDMI

on long-horizon decision-making compared to SOTA offline RL methods and conditio
nal generative models.
****

Divide and Conquer Dynamic Programming: An Almost Linear Time Change Point Detec
tion Methodology in High Dimensions
Wanshan Li, Daren Wang, Alessandro Rinaldo
We develop a novel, general and computationally efficient framework, called Divi
de and Conquer Dynamic Programming (DCDP), for localizing change points in time
series data with high-dimensional features. DCDP deploys a class of greedy algor
ithms that are applicable to a broad variety of high-dimensional statistical mod
els and can enjoy almost linear computational complexity. We investigate the per
formance of DCDP in three commonly studied change point settings in high dimensi
ons: the mean model, the Gaussian graphical model, and the linear regression mod
el. In all three cases, we derive non-asymptotic bounds for the accuracy of the
DCDP change point estimators. We demonstrate that the DCDP procedures consistent
ly estimate the change points with sharp, and in some cases, optimal rates while
 incurring significantly smaller computational costs than the best available alg
orithms. Our findings are supported by extensive numerical experiments on both s
ynthetic and real data.
****

Architecture-Agnostic Masked Image Modeling -- From ViT back to CNN
Siyuan Li, Di Wu, Fang Wu, Zelin Zang, Stan Z. Li
Masked image modeling, an emerging self-supervised pre-training method, has show
n impressive success across numerous downstream vision tasks with Vision transfo
rmers. Its underlying idea is simple: a portion of the input image is masked out
 and then reconstructed via a pre-text task. However, the working principle behi
nd MIM is not well explained, and previous studies insist that MIM primarily wor
ks for the Transformer family but is incompatible with CNNs. In this work, we ob
serve that MIM essentially teaches the model to learn better middle-order intera
ctions among patches for more generalized feature extraction. We then propose an
 Architecture-Agnostic Masked Image Modeling framework (A^{2}MIM), which is comp
atible with both Transformers and CNNs in a unified way. Extensive experiments o
n popular benchmarks show that A^{2}MIM learns better representations without ex
plicit design and endows the backbone model with the stronger capability to tran
sfer to various downstream tasks.
****

Learning Antidote Data to Individual Unfairness
Peizhao Li, Ethan Xia, Hongfu Liu
Fairness is essential for machine learning systems deployed in high-stake applic
ations. Among all fairness notions, individual fairness, deriving from a consens
us that ‘similar individuals should be treated similarly,’ is a vital notion to
describe fair treatment for individual cases. Previous studies typically charact
erize individual fairness as a prediction-invariant problem when perturbing sens
itive attributes on samples, and solve it by Distributionally Robust Optimizatio
n (DRO) paradigm. However, such adversarial perturbations along a direction cove
ring sensitive information used in DRO do not consider the inherent feature corr
elations or innate data constraints, therefore could mislead the model to optimi
ze at off-manifold and unrealistic samples. In light of this drawback, in this p
aper, we propose to learn and generate antidote data that approximately follows
the data distribution to remedy individual unfairness. These generated on-manifo
ld antidote data can be used through a generic optimization procedure along with
 original training data, resulting in a pure pre-processing approach to individu
al unfairness, or can also fit well with the in-processing DRO paradigm. Through
 extensive experiments on multiple tabular datasets, we demonstrate our method r
esists individual unfairness at a minimal or zero cost to predictive utility com
pared to baselines.
****

Propensity Matters: Measuring and Enhancing Balancing for Recommendation
Haoxuan Li, Yanghao Xiao, Chunyuan Zheng, Peng Wu, Peng Cui
Propensity-based weighting methods have been widely studied and demonstrated com

petitive performance in debiased recommendations. Nevertheless, there are still many questions to be addressed. How to estimate the propensity more conducive to debiasing performance? Which metric is more reasonable to measure the quality of the learned propensities? Is it better to make the cross-entropy loss as small as possible when learning propensities? In this paper, we first discuss the potential problems of the previously widely adopted metrics for learned propensities, and propose balanced-mean-squared-error (BMSE) metric for debiased recommendations. Based on BMSE, we propose IPS-V2 and DR-V2 as the estimators of unbiased loss, and theoretically show that IPS-V2 and DR-V2 have greater propensity balancing and smaller variance without sacrificing additional bias. We further propose a co-training method for learning balanced representation and unbiased prediction. Extensive experiments are conducted on three real-world datasets including a large industrial dataset, and the results show that our approach boosts the balancing property and results in enhanced debiasing performance.

****

GraphCleaner: Detecting Mislabelled Samples in Popular Graph Learning Benchmarks

Yuwen Li, Miao Xiong, Bryan Hooi

Label errors have been found to be prevalent in popular text, vision, and audio datasets, which heavily influence the safe development and evaluation of machine learning algorithms. Despite increasing efforts towards improving the quality of generic data types, such as images and texts, the problem of mislabel detection in graph data remains underexplored. To bridge the gap, we explore mislabelling issues in popular real-world graph datasets and propose GraphCleaner, a post-hoc method to detect and correct these mislabelled nodes in graph datasets. GraphCleaner combines the novel ideas of 1) Synthetic Mislabel Dataset Generation, which seeks to generate realistic mislabels; and 2) Neighborhood-Aware Mislabel Detection, where neighborhood dependency is exploited in both labels and base classifier predictions. Empirical evaluations on 6 datasets and 6 experimental settings demonstrate that GraphCleaner outperforms the closest baseline, with an average improvement of 0.14 in F1 score, and 0.16 in MCC. On real-data case studies, GraphCleaner detects real and previously unknown mislabels in popular graph benchmarks: PubMed, Cora, CiteSeer and OGB-arxiv; we find that at least 6.91% of PubMed data is mislabelled or ambiguous, and simply removing these mislabelled data can boost evaluation performance from 86.71% to 89.11%.

****

SMURF-THP: Score Matching-based UnceRtainty quantiFication for Transformer Hawkes Process

Zichong Li, Yanbo Xu, Simiao Zuo, Haoming Jiang, Chao Zhang, Tuo Zhao, Hongyuan Zha

Transformer Hawkes process models have shown to be successful in modeling event sequence data. However, most of the existing training methods rely on maximizing the likelihood of event sequences, which involves calculating some intractable integral. Moreover, the existing methods fail to provide uncertainty quantification for model predictions, e.g., confidence interval for the predicted event’s arrival time. To address these issues, we propose SMURF-THP, a score-based method for learning Transformer Hawkes process and quantifying prediction uncertainty. Specifically, SMURF-THP learns the score function of the event’s arrival time based on a score-matching objective that avoids the intractable computation. With such a learnt score function, we can sample arrival time of events from the predictive distribution. This naturally allows for the quantification of uncertainty by computing confidence intervals over the generated samples. We conduct extensive experiments in both event type prediction and uncertainty quantification on time of arrival. In all the experiments, SMURF-THP outperforms existing likelihood-based methods in confidence calibration while exhibiting comparable prediction accuracy.

****

Horizon-free Learning for Markov Decision Processes and Games: Stochastically Bounded Rewards and Improved Bounds

Shengshi Li, Lin Yang

Horizon dependence is an important difference between reinforcement learning and

other machine learning paradigms. Yet, existing results tackling the (exact) horizon dependence either assume that the reward is bounded per step, introducing unfair comparison, or assume strict total boundedness that requires the sum of rewards to be bounded almost surely – allowing only restricted noise on the reward observation. This paper addresses these limitations by introducing a new relaxation – expected boundedness on rewards, where we allow the reward to be stochastic with only boundedness on the expected sum – opening the door to study horizon-dependence with a much broader set of reward functions with noises. We establish a novel generic algorithm that achieves no-horizon dependence in terms of sample complexity for both Markov Decision Processes (MDP) and Games, via reduction to a good-conditioned auxiliary Markovian environment, in which only “important” state-action pairs are preserved. The algorithm takes only $\tilde{O}(\frac{S^2A}{\epsilon^2})$ episodes interacting with such an environment to achieve an ϵ-optimal policy/strategy (with high probability), improving (zhang, 2022) (which only applies to MDPs with deterministic rewards). Here S is the number of states and A is the number of actions, and the bound is independent of the horizon H.

****

Transcendental Idealism of Planner: Evaluating Perception from Planning Perspective for Autonomous Driving

Weixin Li, Xiaodong Yang

Evaluating the performance of perception modules in autonomous driving is one of the most critical tasks in developing the complex intelligent system. While module-level unit test metrics adopted from traditional computer vision tasks are feasible to some extent, it remains far less explored to measure the impact of perceptual noise on the driving quality of autonomous vehicles in a consistent and holistic manner. In this work, we propose a principled framework that provides a coherent and systematic understanding of the impact an error in the perception module imposes on an autonomous agent’s planning that actually controls the vehicle. Specifically, the planning process is formulated as expected utility maximisation, where all input signals from upstream modules jointly provide a world state description, and the planner strives for the optimal action by maximising the expected utility determined by both world states and actions. We show that, under practical conditions, the objective function can be represented as an inner product between the world state description and the utility function in a Hilbert space. This geometric interpretation enables a novel way to analyse the impact of noise in world state estimation on planning and leads to a universal metric for evaluating perception. The whole framework resembles the idea of transcendental idealism in the classical philosophical literature, which gives the name to our approach.

****

Learning for Edge-Weighted Online Bipartite Matching with Robustness Guarantees

Pengfei Li, Jianyi Yang, Shaolei Ren

Many problems, such as online ad display, can be formulated as online bipartite matching. The crucial challenge lies in the nature of sequentially-revealed online item information, based on which we make irreversible matching decisions at each step. While numerous expert online algorithms have been proposed with bounded worst-case competitive ratios, they may not offer satisfactory performance in average cases. On the other hand, reinforcement learning (RL) has been applied to improve the average performance, but it lacks robustness and can perform arbitrarily poorly. In this paper, we propose a novel RL-based approach to edge-weighted online bipartite matching with robustness guarantees (LOMAR), achieving both good average-case and worst-case performance. The key novelty of LOMAR is a new online switching operation which, based on a judicious condition to hedge against future uncertainties, decides whether to follow the expert’s decision or the RL decision for each online item. We prove that for any $\rho\in[0,1]$, LOMAR is ρ-competitive against any given expert online algorithm. To improve the average performance, we train the RL policy by explicitly considering the online switching operation. Finally, we run empirical experiments to demonstrate the advantages of LOMAR compared to existing baselines.

****
FedVS: Straggler-Resilient and Privacy-Preserving Vertical Federated Learning fo r Split Models
Songze Li, Duanyi Yao, Jin Liu
In a vertical federated learning (VFL) system consisting of a central server and many distributed clients, the training data are vertically partitioned such tha t different features are privately stored on different clients. The problem of s plit VFL is to train a model split between the server and the clients. This pape r aims to address two major challenges in split VFL: 1) performance degradation due to straggling clients during training; and 2) data and model privacy leakage from clients' uploaded data embeddings. We propose FedVS to simultaneously addr ess these two challenges. The key idea of FedVS is to design secret sharing sche mes for the local data and models, such that information-theoretical privacy aga inst colluding clients and curious server is guaranteed, and the aggregation of all clients' embeddings is reconstructed losslessly, via decrypting computation shares from the non-straggling clients. Extensive experiments on various types o f VFL datasets (including tabular, CV, and multi-view) demonstrate the universal advantages of FedVS in straggler mitigation and privacy protection over baselin e protocols.
****
Achieving Hierarchy-Free Approximation for Bilevel Programs with Equilibrium Con straints
Jiayang Li, Jing Yu, Boyi Liu, Yu Nie, Zhaoran Wang
In this paper, we develop an approximation scheme for solving bilevel programs w ith equilibrium constraints, which are generally difficult to solve. Among other things, calculating the first-order derivative in such a problem requires diffe rentiation across the hierarchy, which is computationally intensive, if not proh ibitive. To bypass the hierarchy, we propose to bound such bilevel programs, equ ivalent to multiple-followers Stackelberg games, with two new hierarchy-free pro blems: a T-step Cournot game and a T-step monopoly model. Since they are sta ndard equilibrium or optimization problems, both can be efficiently solved via f irst-order methods. Importantly, we show that the bounds provided by these probl ems — the upper bound by the T-step Cournot game and the lower bound by the $T $-step monopoly model — can be made arbitrarily tight by increasing the step par ameter T for a wide range of problems. We prove that a small T usually suffi ces under appropriate conditions to reach an approximation acceptable for most p ractical purposes. Eventually, the analytical insights are highlighted through n umerical examples.
****
LoSparse: Structured Compression of Large Language Models based on Low-Rank and Sparse Approximation
Yixiao Li, Yifan Yu, Qingru Zhang, Chen Liang, Pengcheng He, Weizhu Chen, Tuo Zh ao
Transformer models have achieved remarkable results in various natural language tasks, but they are often prohibitively large, requiring massive memories and co mputational resources. To re- duce the size and complexity of these models, we p ropose LoSparse (Low-Rank and Sparse ap- proximation), a novel model compression tech- nique that approximates a weight matrix by the sum of a low-rank matrix a nd a sparse matrix. Our method combines the advantages of both low- rank approxi mations and pruning, while avoid- ing their limitations. Low-rank approximation compresses the coherent and expressive parts in neurons, while pruning removes t he incoherent and non-expressive parts in neurons. Pruning enhances the diversit y of low-rank approxima- tions, and low-rank approximation prevents prun- ing fr om losing too many expressive neurons. We evaluate our method on natural languag e under- standing, question answering, and natural lan- guage generation tasks. We show that it signif- icantly outperforms existing compression meth- ods. Our code is publicly available at https: //github.com/yxli2123/LoSparse
****
Nesterov Meets Optimism: Rate-Optimal Separable Minimax Optimization
Chris Junchi Li, Huizhuo Yuan, Gauthier Gidel, Quanquan Gu, Michael Jordan

We propose a new first-order optimization algorithm – AcceleratedGradient-OptimisticGradient (AG-OG) Descent Ascent—for separable convex-concave minimax optimization. The main idea of our algorithm is to carefully leverage the structure of the minimax problem, performing Nesterov acceleration on the individual component and optimistic gradient on the coupling component. Equipped with proper restarting, we show that AG-OG achieves the optimal convergence rate (up to a constant) for a variety of settings, including bilinearly coupled strongly convex-strongly concave minimax optimization (bi-SC-SC), bilinearly coupled convex-strongly concave minimax optimization (bi-C-SC), and bilinear games. We also extend our algorithm to the stochastic setting and achieve the optimal convergence rate in both bi-SC-SC and bi-C-SC settings. AG-OG is the first single-call algorithm with optimal convergence rates in both deterministic and stochastic settings for bilinearly coupled minimax optimization problems.

****

Alternating Local Enumeration (TnALE): Solving Tensor Network Structure Search with Fewer Evaluations

Chao Li, Junhua Zeng, Chunmei Li, Cesar F Caiafa, Qibin Zhao

Tensor network (TN) is a powerful framework in machine learning, but selecting a good TN model, known as TN structure search (TN-SS), is a challenging and computationally intensive task. The recent approach TNLS (Li et al., 2022) showed promising results for this task. However, its computational efficiency is still unaffordable, requiring too many evaluations of the objective function. We propose TnALE, a surprisingly simple algorithm that updates each structure-related variable alternately by local enumeration, greatly reducing the number of evaluations compared to TNLS. We theoretically investigate the descent steps for TNLS and TnALE, proving that both the algorithms can achieve linear convergence up to a constant if a sufficient reduction of the objective is reached in each neighborhood. We further compare the evaluation efficiency of TNLS and TnALE, revealing that $\Omega(2^K)$ evaluations are typically required in TNLS for reaching the objective reduction, while ideally $O(KR)$ evaluations are sufficient in TnALE, where K denotes the dimension of search space and R reflects the “low-rankness” of the neighborhood. Experimental results verify that TnALE can find practically good TN structures with vastly fewer evaluations than the state-of-the-art algorithms.

****

Understanding the Complexity Gains of Single-Task RL with a Curriculum

Qiyang Li, Yuexiang Zhai, Yi Ma, Sergey Levine

Reinforcement learning (RL) problems can be challenging without well-shaped rewards. Prior work on provably efficient RL methods generally proposes to address this issue with dedicated exploration strategies. However, another way to tackle this challenge is to reformulate it as a multi-task RL problem, where the task space contains not only the challenging task of interest but also easier tasks that implicitly function as a curriculum. Such a reformulation opens up the possibility of running existing multi-task RL methods as a more efficient alternative to solving a single challenging task from scratch. In this work, we provide a theoretical framework that reformulates a single-task RL problem as a multi-task RL problem defined by a curriculum. Under mild regularity conditions on the curriculum, we show that sequentially solving each task in the multi-task RL problem is more computationally efficient than solving the original single-task problem, without any explicit exploration bonuses or other exploration strategies. We also show that our theoretical insights can be translated into an effective practical learning algorithm that can accelerate curriculum learning on simulated robotic tasks.

****

Does a Neural Network Really Encode Symbolic Concepts?

Mingjie Li, Quanshi Zhang

Recently, a series of studies have tried to extract interactions between input variables modeled by a DNN and define such interactions as concepts encoded by the DNN. However, strictly speaking, there still lacks a solid guarantee whether such interactions indeed represent meaningful concepts. Therefore, in this paper,

we examine the trustworthiness of interaction concepts from four perspectives. Extensive empirical studies have verified that a well-trained DNN usually encodes sparse, transferable, and discriminative concepts, which is partially aligned with human intuition. The code is released at https://github.com/sjtu-xai-lab/interaction-concept.

****

Cooperative Open-ended Learning Framework for Zero-Shot Coordination

Yang Li, Shao Zhang, Jichen Sun, Yali Du, Ying Wen, Xinbing Wang, Wei Pan

Zero-shot coordination in cooperative artificial intelligence (AI) remains a significant challenge, which means effectively coordinating with a wide range of unseen partners. Previous algorithms have attempted to address this challenge by optimizing fixed objectives within a population to improve strategy or behaviour diversity. However, these approaches can result in a loss of learning and an inability to cooperate with certain strategies within the population, known as cooperative incompatibility. To address this issue, we propose the Cooperative Open-ended LEarning (COLE) framework, which constructs open-ended objectives in cooperative games with two players from the perspective of graph theory to assess and identify the cooperative ability of each strategy. We further specify the framework and propose a practical algorithm that leverages knowledge from game theory and graph theory. Furthermore, an analysis of the learning process of the algorithm shows that it can efficiently overcome cooperative incompatibility. The experimental results in the Overcooked game environment demonstrate that our method outperforms current state-of-the-art methods when coordinating with different-level partners. Our demo is available at https://sites.google.com/view/cole-2023.

****

Offline Reinforcement Learning with Closed-Form Policy Improvement Operators

Jiachen Li, Edwin Zhang, Ming Yin, Qinxun Bai, Yu-Xiang Wang, William Yang Wang

Behavior constrained policy optimization has been demonstrated to be a successful paradigm for tackling Offline Reinforcement Learning. By exploiting historical transitions, a policy is trained to maximize a learned value function while constrained by the behavior policy to avoid a significant distributional shift. In this paper, we propose our closed-form policy improvement operators. We make a novel observation that the behavior constraint naturally motivates the use of first-order Taylor approximation, leading to a linear approximation of the policy objective. Additionally, as practical datasets are usually collected by heterogeneous policies, we model the behavior policies as a Gaussian Mixture and overcome the induced optimization difficulties by leveraging the LogSumExp’s lower bound and Jensen’s Inequality, giving rise to a closed-form policy improvement operator. We instantiate both one-step and iterative offline RL algorithms with our novel policy improvement operators and empirically demonstrate their effectiveness over state-of-the-art algorithms on the standard D4RL benchmark. Our code is available at https://cfpi-icml23.github.io/.

****

Optimal Arms Identification with Knapsacks

Shaoang Li, Lan Zhang, Yingqi Yu, Xiangyang Li

Best Arm Identification (BAI) is a general online pure exploration framework to identify optimal decisions among candidates via sequential interactions. We pioneer the Optimal Arms identification with Knapsacks (OAK) problem, which extends the BAI setting to model the resource consumption. We present a novel OAK algorithm and prove the upper bound of our algorithm by exploring the relationship between selecting optimal actions and the structure of the feasible region. Our analysis introduces a new complexity measure, which builds a bridge between the OAK setting and bandits with knapsacks problem. We establish the instance-dependent lower bound for the OAK problem based on the new complexity measure. Our results show that the proposed algorithm achieves a near-optimal probability bound for the OAK problem. In addition, we demonstrate that our algorithm recovers or improves the state-of-the-art upper bounds for several special cases, including the simple OAK setting and some classical pure exploration problems.

****

Internally Rewarded Reinforcement Learning

Mengdi Li, Xufeng Zhao, Jae Hee Lee, Cornelius Weber, Stefan Wermter
We study a class of reinforcement learning problems where the reward signals for policy learning are generated by a discriminator that is dependent on and jointly optimized with the policy. This interdependence between the policy and the discriminator leads to an unstable learning process because reward signals from an immature discriminator are noisy and impede policy learning, and conversely, an under-optimized policy impedes discriminator learning. We call this learning setting $\textit{Internally Rewarded Reinforcement Learning}$ (IRRL) as the reward is not provided directly by the environment but $\textit{internally}$ by the discriminator. In this paper, we formally formulate IRRL and present a class of problems that belong to IRRL. We theoretically derive and empirically analyze the effect of the reward function in IRRL and based on these analyses propose the clipped linear reward function. Experimental results show that the proposed reward function can consistently stabilize the training process by reducing the impact of reward noise, which leads to faster convergence and higher performance compared with baselines in diverse tasks.
****

Trustworthy Policy Learning under the Counterfactual No-Harm Criterion
Haoxuan Li, Chunyuan Zheng, Yixiao Cao, Zhi Geng, Yue Liu, Peng Wu
Trustworthy policy learning has significant importance in making reliable and harmless treatment decisions for individuals. Previous policy learning approaches aim at the well-being of subgroups by maximizing the utility function (e.g., conditional average causal effects, post-view click-through&conversion rate in recommendations), however, individual-level counterfactual no-harm criterion has rarely been discussed. In this paper, we first formalize the counterfactual no-harm criterion for policy learning from a principal stratification perspective. Next, we propose a novel upper bound for the fraction negatively affected by the policy and show the consistency and asymptotic normality of the estimator. Based on the estimators for the policy utility and harm upper bounds, we further propose a policy learning approach that satisfies the counterfactual no-harm criterion, and prove its consistency to the optimal policy reward for parametric and non-parametric policy classes, respectively. Extensive experiments are conducted to show the effectiveness of the proposed policy learning approach for satisfying the counterfactual no-harm criterion.
****

Structured Cooperative Learning with Graphical Model Priors
Shuangtong Li, Tianyi Zhou, Xinmei Tian, Dacheng Tao
We study how to train personalized models for different tasks on decentralized devices with limited local data. We propose "Structured Cooperative Learning (SCooL)", in which a cooperation graph across devices is generated by a graphical model prior to automatically coordinate mutual learning between devices. By choosing graphical models enforcing different structures, we can derive a rich class of existing and novel decentralized learning algorithms via variational inference. In particular, we show three instantiations of SCooL that adopt Dirac distribution, stochastic block model (SBM), and attention as the prior generating cooperation graphs. These EM-type algorithms alternate between updating the cooperation graph and cooperative learning of local models. They can automatically capture the cross-task correlations among devices by only monitoring their model updating in order to optimize the cooperation graph. We evaluate SCooL and compare it with existing decentralized learning methods on an extensive set of benchmarks, on which SCooL always achieves the highest accuracy of personalized models and significantly outperforms other baselines on communication efficiency. Our code is available at https://github.com/ShuangtongLi/SCooL.
****

Low Complexity Homeomorphic Projection to Ensure Neural-Network Solution Feasibility for Optimization over (Non-)Convex Set
Enming Liang, Minghua Chen, Steven Low
There has been growing interest in employing neural network (NN) to directly solve constrained optimization problems with low run-time complexity. However, it is non-trivial to ensure NN solutions strictly satisfying problem constraints due

to inherent NN prediction errors. Existing feasibility-ensuring methods either are computationally expensive or lack performance guarantee. In this paper, we propose homeomorphic projection as a low-complexity scheme to guarantee NN solution feasibility for optimization over a general set homeomorphic to a unit ball, covering all compact convex sets and certain classes of nonconvex sets. The idea is to (i) learn a minimum distortion homeomorphic mapping between the constraint set and a unit ball using an invertible NN (INN), and then (ii) perform a simple bisection operation concerning the unit ball so that the INN-mapped final solution is feasible with respect to the constraint set with minor distortion-induced optimality loss. We prove the feasibility guarantee and bound the optimality loss under mild conditions. Simulation results, including those for non-convex AC-OPF problems in power grid operation, show that homeomorphic projection outperforms existing methods in solution feasibility and run-time complexity, while achieving similar optimality loss.

****

Consistency of Multiple Kernel Clustering

Weixuan Liang, Xinwang Liu, Yong Liu, Chuan Ma, Yunping Zhao, Zhe Liu, En Zhu

Consistency plays an important role in learning theory. However, in multiple kernel clustering (MKC), the consistency of kernel weights has not been sufficiently investigated. In this work, we fill this gap with a non-asymptotic analysis on the consistency of kernel weights of a novel method termed SimpleMKKM. Under the assumptions of the eigenvalue gap, we give an infinity norm bound as $\widetilde{\mathcal{O}}(k/\sqrt{n})$, where k is the number of clusters and n is the number of samples. On this basis, we establish an upper bound for the excess clustering risk. Moreover, we study the difference of the kernel weights learned from n samples and r points sampled without replacement, and derive its upper bound as $\widetilde{\mathcal{O}}(k\cdot\sqrt{1/r-1/n})$. Based on the above results, we propose a novel strategy with Nyström method to enable SimpleMKKM to handle large-scale datasets with a theoretical learning guarantee. Finally, extensive experiments are conducted to verify the theoretical results and the effectiveness of the proposed large-scale strategy.

****

A Distribution Optimization Framework for Confidence Bounds of Risk Measures

Hao Liang, Zhi-Quan Luo

We present a distribution optimization framework that significantly improves confidence bounds for various risk measures compared to previous methods. Our framework encompasses popular risk measures such as the entropic risk measure, conditional value at risk (CVaR), spectral risk measure, distortion risk measure, equivalent certainty, and rank-dependent expected utility, which are well established in risk-sensitive decision-making literature. To achieve this, we introduce two estimation schemes based on concentration bounds derived from the empirical distribution, specifically using either the Wasserstein distance or the supremum distance. Unlike traditional approaches that add or subtract a confidence radius from the empirical risk measures, our proposed schemes evaluate a specific transformation of the empirical distribution based on the distance. Consequently, our confidence bounds consistently yield tighter results compared to previous methods. We further verify the efficacy of the proposed framework by providing tighter problem-dependent regret bound for the CVaR bandit.

****

Accuracy on the Curve: On the Nonlinear Correlation of ML Performance Between Data Subpopulations

Weixin Liang, Yining Mao, Yongchan Kwon, Xinyu Yang, James Zou

Understanding the performance of machine learning (ML) models across diverse data distributions is critically important for reliable applications. Despite recent empirical studies positing a near-perfect linear correlation between in-distribution (ID) and out-of-distribution (OOD) accuracies, we empirically demonstrate that this correlation is more nuanced under subpopulation shifts. Through rigorous experimentation and analysis across a variety of datasets, models, and training epochs, we demonstrate that OOD performance often has a nonlinear correlation with ID performance in subpopulation shifts. Our findings, which contrast prev

ious studies that have posited a linear correlation in model performance during distribution shifts, reveal a "moon shape" correlation (parabolic uptrend curve) between the test performance on the majority subpopulation and the minority subpopulation. This non-trivial nonlinear correlation holds across model architectures, hyperparameters, training durations, and the imbalance between subpopulations. Furthermore, we found that the nonlinearity of this "moon shape" is causally influenced by the degree of spurious correlations in the training data. Our controlled experiments show that stronger spurious correlation in the training data creates more nonlinear performance correlation. We provide complementary experimental and theoretical analyses for this phenomenon, and discuss its implications for ML reliability and fairness. Our work highlights the importance of understanding the nonlinear effects of model improvement on performance in different subpopulations, and has the potential to inform the development of more equitable and responsible machine learning models.

****

AdaptDiffuser: Diffusion Models as Adaptive Self-evolving Planners

Zhixuan Liang, Yao Mu, Mingyu Ding, Fei Ni, Masayoshi Tomizuka, Ping Luo

Diffusion models have demonstrated their powerful generative capability in many tasks, with great potential to serve as a paradigm for offline reinforcement learning. However, the quality of the diffusion model is limited by the insufficient diversity of training data, which hinders the performance of planning and the generalizability to new tasks. This paper introduces AdaptDiffuser, an evolutionary planning method with diffusion that can self-evolve to improve the diffusion model hence a better planner, not only for seen tasks but can also adapt to unseen tasks. AdaptDiffuser enables the generation of rich synthetic expert data for goal-conditioned tasks using guidance from reward gradients. It then selects high-quality data via a discriminator to finetune the diffusion model, which improves the generalization ability to unseen tasks. Empirical experiments on two benchmark environments and two carefully designed unseen tasks in KUKA industrial robot arm and Maze2D environments demonstrate the effectiveness of AdaptDiffuser. For example, AdaptDiffuser not only outperforms the previous art Diffuser by 20.8% on Maze2D and 7.5% on MuJoCo locomotion, but also adapts better to new tasks, e.g., KUKA pick-and-place, by 27.9% without requiring additional expert data. More visualization results and demo videos could be found on our project page.

****

Learning Compiler Pass Orders using Coreset and Normalized Value Prediction

Youwei Liang, Kevin Stone, Ali Shameli, Chris Cummins, Mostafa Elhoushi, Jiadong Guo, Benoit Steiner, Xiaomeng Yang, Pengtao Xie, Hugh James Leather, Yuandong Tian

Finding the optimal pass sequence of compilation can lead to a significant reduction in program size. Prior works on compilation pass ordering have two major drawbacks. They either require an excessive budget (in terms of the number of compilation passes) at compile time or fail to generalize to unseen programs. In this work, instead of predicting passes sequentially, we directly learn a policy on the pass sequence space, which outperforms the default -Oz flag by an average of 4.5% over a large collection (4683) of unseen code repositories from diverse domains across 14 datasets. To achieve this, we first identify a small set (termed coreset) of pass sequences that generally optimize the size of most programs. Then, a policy is learned to pick the optimal sequences by predicting the normalized values of the pass sequences in the coreset. Our results demonstrate that existing human-designed compiler passes can be improved with a simple yet effective technique that leverages pass sequence space which contains dense rewards, while approaches operating on the individual pass space may suffer from issues of sparse reward, and do not generalize well to held-out programs from different domains. Website: https://rlcompopt.github.io.

****

Adversarial Example Does Good: Preventing Painting Imitation from Diffusion Models via Adversarial Examples

Chumeng Liang, Xiaoyu Wu, Yang Hua, Jiaru Zhang, Yiming Xue, Tao Song, Zhengui Xue, Ruhui Ma, Haibing Guan

Recently, Diffusion Models (DMs) boost a wave in AI for Art yet raise new copyri
ght concerns, where infringers benefit from using unauthorized paintings to trai
n DMs and generate novel paintings in a similar style. To address these emerging
 copyright violations, in this paper, we are the first to explore and propose to
 utilize adversarial examples for DMs to protect human-created artworks. Specifi
cally, we first build a theoretical framework to define and evaluate the adversa
rial examples for DMs. Then, based on this framework, we design a novel algorith
m to generate these adversarial examples, named AdvDM, which exploits a Monte-Ca
rlo estimation of adversarial examples for DMs by optimizing upon different late
nt variables sampled from the reverse process of DMs. Extensive experiments show
 that the generated adversarial examples can effectively hinder DMs from extract
ing their features. Therefore, our method can be a powerful tool for human artis
ts to protect their copyright against infringers equipped with DM-based AI-for-A
rt applications. The code of our method is available on GitHub: https://github.c
om/mist-project/mist.git.
****

CLUSTSEG: Clustering for Universal Segmentation
James Chenhao Liang, Tianfei Zhou, Dongfang Liu, Wenguan Wang
We present CLUSTSEG, a general, transformer-based framework that tackles differe
nt image segmentation tasks ($i.e.,$ superpixel, semantic, instance, and panopti
c) through a unified, neural clustering scheme. Regarding queries as cluster cen
ters, CLUSTSEG is innovative in two aspects: 1) cluster centers are initialized
in heterogeneous ways so as to pointedly address task-specific demands ($e.g.,$
instance- or category-level distinctiveness), yet without modifying the architec
ture; and 2) pixel-cluster assignment, formalized in a cross-attention fashion,
is alternated with cluster center update, yet without learning additional parame
ters. These innovations closely link CLUSTSEG to EM clustering and make it a tra
nsparent and powerful framework that yields superior results across the above se
gmentation tasks.
****

Conformal Inference is (almost) Free for Neural Networks Trained with Early Stop
ping
Ziyi Liang, Yanfei Zhou, Matteo Sesia
Early stopping based on hold-out data is a popular regularization technique desi
gned to mitigate overfitting and increase the predictive accuracy of neural netw
orks. Models trained with early stopping often provide relatively accurate predi
ctions, but they generally still lack precise statistical guarantees unless they
 are further calibrated using independent hold-out data. This paper addresses th
e above limitation with conformalized early stopping: a novel method that combin
es early stopping with conformal calibration while efficiently recycling the sam
e hold-out data. This leads to models that are both accurate and able to provide
 exact predictive inferences without multiple data splits nor overly conservativ
e adjustments. Practical implementations are developed for different learning ta
sks—outlier detection, multi-class classification, regression—and their competit
ive performance is demonstrated on real data.
****

Less is More: Task-aware Layer-wise Distillation for Language Model Compression
Chen Liang, Simiao Zuo, Qingru Zhang, Pengcheng He, Weizhu Chen, Tuo Zhao
Layer-wise distillation is a powerful tool to compress large models (i.e. teache
r models) into small ones (i.e., student models). The student distills knowledge
 from the teacher by mimicking the hidden representations of the teacher at ever
y intermediate layer. However, layer-wise distillation is difficult. Since the s
tudent has a smaller model capacity than the teacher, it is often under-fitted.
Furthermore, the hidden representations of the teacher contain redundant informa
tion that the student does not necessarily need for the target task’s learning.
To address these challenges, we propose a novel Task-aware layEr-wise Distillati
on (TED). TED designs task-aware filters to align the hidden representations of
the student and the teacher at each layer. The filters select the knowledge that
 is useful for the target task from the hidden representations. As such, TED red
uces the knowledge gap between the two models and helps the student to fit bette

r on the target task. We evaluate TED in two scenarios: continual pre-training a
nd fine-tuning. TED demonstrates significant and consistent improvements over ex
isting distillation methods in both scenarios. Code is available at https://gith
ub.com/cliang1453/task-aware-distillation.
****
Statistical Inference and A/B Testing for First-Price Pacing Equilibria
Luofeng Liao, Christian Kroer
We initiate the study of statistical inference and A/B testing for first-price p
acing equilibria (FPPE). The FPPE model captures the dynamics resulting from lar
ge-scale first-price auction markets where buyers use pacing-based budget manage
ment. Such markets arise in the context of internet advertising, where budgets a
re prevalent. We propose a statistical framework for the FPPE model, in which a
limit FPPE with a continuum of items models the long-run steady-state behavior o
f the auction platform, and an observable FPPE consisting of a finite number of
items provides the data to estimate primitives of the limit FPPE, such as revenu
e, Nash social welfare (a fair metric of efficiency), and other parameters of in
terest. We develop central limit theorems and asymptotically valid confidence in
tervals. Furthermore, we establish the asymptotic local minimax optimality of ou
r estimators. We then show that the theory can be used for conducting statistica
lly valid A/B testing on auction platforms. Numerical simulations verify our cen
tral limit theorems, and empirical coverage rates for our confidence intervals a
gree with our theory.
****
Supervised Metric Learning to Rank for Retrieval via Contextual Similarity Optim
ization
Christopher Liao, Theodoros Tsiligkaridis, Brian Kulis
There is extensive interest in metric learning methods for image retrieval. Many
 metric learning loss functions focus on learning a correct ranking of training
samples, but strongly overfit semantically inconsistent labels and require a lar
ge amount of data. To address these shortcomings, we propose a new metric learni
ng method, called contextual loss, which optimizes contextual similarity in addi
tion to cosine similarity. Our contextual loss implicitly enforces semantic cons
istency among neighbors while converging to the correct ranking. We empirically
show that the proposed loss is more robust to label noise, and is less prone to
overfitting even when a large portion of train data is withheld. Extensive exper
iments demonstrate that our method achieves a new state-of-the-art across four i
mage retrieval benchmarks and multiple different evaluation settings. Code is av
ailable at: https://github.com/Chris210634/metric-learning-using-contextual-simi
larity
****
Revisiting Domain Randomization via Relaxed State-Adversarial Policy Optimizatio
n
Yun-Hsuan Lien, Ping-Chun Hsieh, Yu-Shuen Wang
Domain randomization (DR) is widely used in reinforcement learning (RL) to bridg
e the gap between simulation and reality by maximizing its average returns under
 the perturbation of environmental parameters. However, even the most complex si
mulators cannot capture all details in reality due to finite domain parameters a
nd simplified physical models. Additionally, the existing methods often assume t
hat the distribution of domain parameters belongs to a specific family of probab
ility functions, such as normal distributions, which may not be correct. To over
come these limitations, we propose a new approach to DR by rethinking it from th
e perspective of adversarial state perturbation, without the need for reconfigur
ing the simulator or relying on prior knowledge about the environment. We also a
ddress the issue of over-conservatism that can occur when perturbing agents to t
he worst states during training by introducing a Relaxed State-Adversarial Algor
ithm that simultaneously maximizes the average-case and worst-case returns. We e
valuate our method by comparing it to state-of-the-art methods, providing experi
mental results and theoretical proofs to verify its effectiveness. Our source co
de and appendix are available at https://github.com/sophialien/RAPPO.
****

Variational Open-Domain Question Answering
Valentin Liévin, Andreas Geert Motzfeldt, Ida Riis Jensen, Ole Winther
Retrieval-augmented models have proven to be effective in natural language processing tasks, yet there remains a lack of research on their optimization using variational inference. We introduce the Variational Open-Domain (VOD) framework for end-to-end training and evaluation of retrieval-augmented models, focusing on open-domain question answering and language modelling. The VOD objective, a self-normalized estimate of the Rényi variational bound, approximates the task marginal likelihood and is evaluated under samples drawn from an auxiliary sampling distribution (cached retriever and/or approximate posterior). It remains tractable, even for retriever distributions defined on large corpora. We demonstrate VOD's versatility by training reader-retriever BERT-sized models on multiple-choice medical exam questions. On the MedMCQA dataset, we outperform the domain-tuned Med-PaLM by +5.3% despite using 2.500$\times$ fewer parameters. Our retrieval-augmented BioLinkBERT model scored 62.9% on the MedMCQA and 55.0% on the MedQA-USMLE. Last, we show the effectiveness of our learned retriever component in the context of medical semantic search.
****
Generating Novel, Designable, and Diverse Protein Structures by Equivariantly Diffusing Oriented Residue Clouds
Yeqing Lin, Mohammed Alquraishi
Proteins power a vast array of functional processes in living cells. The capability to create new proteins with designed structures and functions would thus enable the engineering of cellular behavior and development of protein-based therapeutics and materials. Structure-based protein design aims to find structures that are designable (can be realized by a protein sequence), novel (have dissimilar geometry from natural proteins), and diverse (span a wide range of geometries). While advances in protein structure prediction have made it possible to predict structures of novel protein sequences, the combinatorially large space of sequences and structures limits the practicality of search-based methods. Generative models provide a compelling alternative, by implicitly learning the low-dimensional structure of complex data distributions. Here, we leverage recent advances in denoising diffusion probabilistic models and equivariant neural networks to develop Genie, a generative model of protein structures that performs discrete-time diffusion using a cloud of oriented reference frames in 3D space. Through in silico evaluations, we demonstrate that Genie generates protein backbones that are more designable, novel, and diverse than existing models. This indicates that Genie is capturing key aspects of the distribution of protein structure space and facilitates protein design with high success rates. Code for generating new proteins and training new versions of Genie is available at https://github.com/aqlaboratory/genie.
****
Hyperbolic Diffusion Embedding and Distance for Hierarchical Representation Learning
Ya-Wei Eileen Lin, Ronald R. Coifman, Gal Mishne, Ronen Talmon
Finding meaningful representations and distances of hierarchical data is important in many fields. This paper presents a new method for hierarchical data embedding and distance. Our method relies on combining diffusion geometry, a central approach to manifold learning, and hyperbolic geometry. Specifically, using diffusion geometry, we build multi-scale densities on the data, aimed to reveal their hierarchical structure, and then embed them into a product of hyperbolic spaces. We show theoretically that our embedding and distance recover the underlying hierarchical structure. In addition, we demonstrate the efficacy of the proposed method and its advantages compared to existing methods on graph embedding benchmarks and hierarchical datasets.
****
Simplifying Momentum-based Positive-definite Submanifold Optimization with Applications to Deep Learning
Wu Lin, Valentin Duruisseaux, Melvin Leok, Frank Nielsen, Mohammad Emtiyaz Khan, Mark Schmidt

Riemannian submanifold optimization with momentum is computationally challenging because, to ensure that the iterates remain on the submanifold, we often need to solve difficult differential equations. Here, we simplify such difficulties for a class of structured symmetric positive-definite matrices with the affine-invariant metric. We do so by proposing a generalized version of the Riemannian normal coordinates that dynamically orthonormalizes the metric and locally converts the problem into an unconstrained problem in the Euclidean space. We use our approach to simplify existing approaches for structured covariances and develop matrix-inverse-free 2^nd-order optimizers for deep learning in low precision settings.
****
Text Generation with Diffusion Language Models: A Pre-training Approach with Continuous Paragraph Denoise
Zhenghao Lin, Yeyun Gong, Yelong Shen, Tong Wu, Zhihao Fan, Chen Lin, Nan Duan, Weizhu Chen
In this paper, we introduce a novel dIffusion language modEl pre-training framework for text generation, which we call GENIE. GENIE is a large-scale pre-trained diffusion language model that consists of an encoder and a diffusion-based decoder, which can generate text by gradually transforming a random noise sequence into a coherent text sequence. To pre-train GENIE on a large-scale language corpus, we design a new continuous paragraph denoise objective, which encourages the diffusion-decoder to reconstruct a clean text paragraph from a corrupted version, while preserving the semantic and syntactic coherence. We evaluate GENIE on four downstream text generation benchmarks, namely XSum, CNN/DailyMail, Gigaword, and CommonGen. Our experimental results show that GENIE achieves comparable performance with the state-of-the-art autoregressive models on these benchmarks, and generates more diverse text samples. The code and models of GENIE are available at https://github.com/microsoft/ProphetNet/tree/master/GENIE.
****
Self-supervised Neural Factor Analysis for Disentangling Utterance-level Speech Representations
Weiwei Lin, Chenhang He, Man-Wai Mak, Youzhi Tu
Self-supervised learning (SSL) speech models such as wav2vec and HuBERT have demonstrated state-of-the-art performance on automatic speech recognition (ASR) and proved to be extremely useful in low label-resource settings. However, the success of SSL models has yet to transfer to utterance-level tasks such as speaker, emotion, and language recognition, which still require supervised fine-tuning of the SSL models to obtain good performance. We argue that the problem is caused by the lack of disentangled representations and an utterance-level learning objective for these tasks. Inspired by how HuBERT uses clustering to discover hidden acoustic units, we formulate a factor analysis (FA) model that uses the discovered hidden acoustic units to align the SSL features. The underlying utterance-level representations are disentangled using probabilistic inference on the aligned features. Furthermore, the variational lower bound derived from the FA model provides an utterance-level objective, allowing error gradients to be backpropagated to the Transformer layers to learn highly discriminative acoustic units. When used in conjunction with HuBERT’s masked prediction training, our models outperform the current best model, WavLM, on all utterance-level non-semantic tasks on the SUPERB benchmark with only 20% of labeled data.
****
Theory on Forgetting and Generalization of Continual Learning
Sen Lin, Peizhong Ju, Yingbin Liang, Ness Shroff
Continual learning (CL), which aims to learn a sequence of tasks, has attracted significant recent attention. However, most work has focused on the experimental performance of CL, and theoretical studies of CL are still limited. In particular, there is a lack of understanding on what factors are important and how they affect "catastrophic forgetting" and generalization performance. To fill this gap, our theoretical analysis, under overparameterized linear models, provides the first-known explicit form of the expected forgetting and generalization error for a general CL setup with an arbitrary number of tasks. Further analysis of suc

h a key result yields a number of theoretical explanations about how overparamet erization, task similarity, and task ordering affect both forgetting and general ization error of CL. More interestingly, by conducting experiments on real datas ets using deep neural networks (DNNs), we show that some of these insights even go beyond the linear models and can be carried over to practical setups. In part icular, we use concrete examples to show that our results not only explain some interesting empirical observations in recent studies, but also motivate better p ractical algorithm designs of CL.
****

Accelerated Cyclic Coordinate Dual Averaging with Extrapolation for Composite Co nvex Optimization

Cheuk Yin Lin, Chaobing Song, Jelena Diakonikolas

Exploiting partial first-order information in a cyclic way is arguably the most natural strategy to obtain scalable first-order methods. However, despite their wide use in practice, cyclic schemes are far less understood from a theoretical perspective than their randomized counterparts. Motivated by a recent success in analyzing an extrapolated cyclic scheme for generalized variational inequalitie s, we propose an Accelerated Cyclic Coordinate Dual Averaging with Extrapolation (A-CODER) method for composite convex optimization, where the objective functio n can be expressed as the sum of a smooth convex function accessible via a gradi ent oracle and a convex, possibly nonsmooth, function accessible via a proximal oracle. We show that A-CODER attains the optimal convergence rate with improved dependence on the number of blocks compared to prior work. Furthermore, for the setting where the smooth component of the objective function is expressible in a finite sum form, we introduce a variance-reduced variant of A-CODER, VR-A-CODER , with state-of-the-art complexity guarantees. Finally, we demonstrate the effec tiveness of our algorithms through numerical experiments.
****

Safe Offline Reinforcement Learning with Real-Time Budget Constraints

Qian Lin, Bo Tang, Zifan Wu, Chao Yu, Shangqin Mao, Qianlong Xie, Xingxing Wang, Dong Wang

Aiming at promoting the safe real-world deployment of Reinforcement Learning (RL), research on safe RL has made significant progress in recent years. However, m ost existing works in the literature still focus on the online setting where ris ky violations of the safety budget are likely to be incurred during training. Be sides, in many realworld applications, the learned policy is required to respond to dynamically determined safety budgets (i.e., constraint threshold) in real t ime. In this paper, we target at the above real-time budget constraint problem u nder the offline setting, and propose Trajectory-based REal-time Budget Inferenc e (TREBI) as a novel solution that approaches this problem from the perspective of trajectory distribution. Theoretically, we prove an error bound of the estima tion on the episodic reward and cost under the offline setting and thus provide a performance guarantee for TREBI. Empirical results on a wide range of simulati on tasks and a real-world large-scale advertising application demonstrate the ca pability of TREBI in solving real-time budget constraint problems under offline settings.
****

Probabilistic Unrolling: Scalable, Inverse-Free Maximum Likelihood Estimation fo r Latent Gaussian Models

Alexander Lin, Bahareh Tolooshams, Yves Atchade, Demba E. Ba

Latent Gaussian models have a rich history in statistics and machine learning, w ith applications ranging from factor analysis to compressed sensing to time seri es analysis. The classical method for maximizing the likelihood of these models is the expectation-maximization (EM) algorithm. For problems with high-dimension al latent variables and large datasets, EM scales poorly because it needs to inv ert as many large covariance matrices as the number of data points. We introduce probabilistic unrolling, a method that combines Monte Carlo sampling with itera tive linear solvers to circumvent matrix inversion. Our theoretical analyses rev eal that unrolling and backpropagation through the iterations of the solver can accelerate gradient estimation for maximum likelihood estimation. In experiments

on simulated and real data, we demonstrate that probabilistic unrolling learns latent Gaussian models up to an order of magnitude faster than gradient EM, with minimal losses in model performance.

****

Fast Online Value-Maximizing Prediction Sets with Conformal Cost Control

Zhen Lin, Shubhendu Trivedi, Cao Xiao, Jimeng Sun

Many real-world multi-label prediction problems involve set-valued predictions that must satisfy specific requirements dictated by downstream usage. We focus on a typical scenario where such requirements, separately encoding value and cost, compete with each other. For instance, a hospital might expect a smart diagnosis system to capture as many severe, often co-morbid, diseases as possible (the value), while maintaining strict control over incorrect predictions (the cost). We present a general pipeline, dubbed as FavMac, to maximize the value while controlling the cost in such scenarios. FavMac can be combined with almost any multi-label classifier, affording distribution-free theoretical guarantees on cost control. Moreover, unlike prior works, FavMac can handle real-world large-scale applications via a carefully designed online update mechanism, which is of independent interest. Our methodological and theoretical contributions are supported by experiments on several healthcare tasks and synthetic datasets - FavMac furnishes higher value compared with several variants and baselines while maintaining strict cost control.

****

Unveiling The Mask of Position-Information Pattern Through the Mist of Image Features

Chieh Hubert Lin, Hung-Yu Tseng, Hsin-Ying Lee, Maneesh Kumar Singh, Ming-Hsuan Yang

Recent studies have shown that paddings in convolutional neural networks encode absolute position information which can negatively affect the model performance for certain tasks. However, existing metrics for quantifying the strength of positional information remain unreliable and frequently lead to erroneous results. To address this issue, we propose novel metrics for measuring and visualizing the encoded positional information. We formally define the encoded information as Position-information Pattern from Padding (PPP) and conduct a series of experiments to study its properties as well as its formation. The proposed metrics measure the presence of positional information more reliably than the existing metrics based on PosENet and tests in F-Conv. We also demonstrate that for any extant (and proposed) padding schemes, PPP is primarily a learning artifact and is less dependent on the characteristics of the underlying padding schemes.

****

Fair yet Asymptotically Equal Collaborative Learning

Xiaoqiang Lin, Xinyi Xu, See-Kiong Ng, Chuan-Sheng Foo, Bryan Kian Hsiang Low

In collaborative learning with streaming data, nodes (e.g., organizations) jointly and continuously learn a machine learning (ML) model by sharing the latest model updates computed from their latest streaming data. For the more resourceful nodes to be willing to share their model updates, they need to be fairly incentivized. This paper explores an incentive design that guarantees fairness so that nodes receive rewards commensurate to their contributions. Our approach leverages an explore-then-exploit formulation to estimate the nodes' contributions (i.e., exploration) for realizing our theoretically guaranteed fair incentives (i.e., exploitation). However, we observe a "rich get richer" phenomenon arising from the existing approaches to guarantee fairness and it discourages the participation of the less resourceful nodes. To remedy this, we additionally preserve asymptotic equality, i.e., less resourceful nodes achieve equal performance eventually to the more resourceful/"rich" nodes. We empirically demonstrate in two settings with real-world streaming data: federated online incremental learning and federated reinforcement learning, that our proposed approach outperforms existing baselines in fairness and learning performance while remaining competitive in preserving equality.

****

Efficient Approximations of Complete Interatomic Potentials for Crystal Property

Prediction
Yuchao Lin, Keqiang Yan, Youzhi Luo, Yi Liu, Xiaoning Qian, Shuiwang Ji
We study property prediction for crystal materials. A crystal structure consists of a minimal unit cell that is repeated infinitely in 3D space. How to accurately represent such repetitive structures in machine learning models remains unresolved. Current methods construct graphs by establishing edges only between nearby nodes, thereby failing to faithfully capture infinite repeating patterns and distant interatomic interactions. In this work, we propose several innovations to overcome these limitations. First, we propose to model physics-principled interatomic potentials directly instead of only using distances as in many existing methods. These potentials include the Coulomb potential, London dispersion potential, and Pauli repulsion potential. Second, we model the complete set of potentials among all atoms, instead of only between nearby atoms as in existing methods. This is enabled by our approximations of infinite potential summations with provable error bounds. We further develop efficient algorithms to compute the approximations. Finally, we propose to incorporate our computations of complete interatomic potentials into message passing neural networks for representation learning. We perform experiments on the JARVIS and Materials Project benchmarks for evaluation. Results show that the use of interatomic potentials and complete interatomic potentials leads to consistent performance improvements with reasonable computational costs. Our code is publicly available as part of the AIRS library (https://github.com/divelab/AIRS).
****
Continuation Path Learning for Homotopy Optimization
Xi Lin, Zhiyuan Yang, Xiaoyuan Zhang, Qingfu Zhang
Homotopy optimization is a traditional method to deal with a complicated optimization problem by solving a sequence of easy-to-hard surrogate subproblems. However, this method can be very sensitive to the continuation schedule design and might lead to a suboptimal solution to the original problem. In addition, the intermediate solutions, often ignored by classic homotopy optimization, could be useful for many real-world applications. In this work, we propose a novel model-based approach to learn the whole continuation path for homotopy optimization, which contains infinite intermediate solutions for any surrogate subproblems. Rather than the classic unidirectional easy-to-hard optimization, our method can simultaneously optimize the original problem and all surrogate subproblems in a collaborative manner. The proposed model also supports the real-time generation of any intermediate solution, which could be desirable for many applications. Experimental studies on different problems show that our proposed method can significantly improve the performance of homotopy optimization and provide extra helpful information to support better decision-making.
****
Speed-Oblivious Online Scheduling: Knowing (Precise) Speeds is not Necessary
Alexander Lindermayr, Nicole Megow, Martin Rapp
We consider online scheduling on unrelated (heterogeneous) machines in a speed-oblivious setting, where an algorithm is unaware of the exact job-dependent processing speeds. We show strong impossibility results for clairvoyant and non-clairvoyant algorithms and overcome them in models inspired by practical settings: (i) we provide competitive learning-augmented algorithms, assuming that (possibly erroneous) predictions on the speeds are given, and (ii) we provide competitive algorithms for the speed-ordered model, where a single global order of machines according to their unknown job-dependent speeds is known. We prove strong theoretical guarantees and evaluate our findings on a representative heterogeneous multi-core processor. These seem to be the first empirical results for scheduling algorithms with predictions that are evaluated in a non-synthetic hardware environment.
****
Graph Mixup with Soft Alignments
Hongyi Ling, Zhimeng Jiang, Meng Liu, Shuiwang Ji, Na Zou
We study graph data augmentation by mixup, which has been used successfully on images. A key operation of mixup is to compute a convex combination of a pair of

inputs. This operation is straightforward for grid-like data, such as images, but challenging for graph data. The key difficulty lies in the fact that different graphs typically have different numbers of nodes, and thus there lacks a node-level correspondence between graphs. In this work, we propose S-Mixup, a simple yet effective mixup method for graph classification by soft alignments. Specifically, given a pair of graphs, we explicitly obtain node-level correspondence via computing a soft assignment matrix to match the nodes between two graphs. Based on the soft assignments, we transform the adjacency and node feature matrices of one graph, so that the transformed graph is aligned with the other graph. In this way, any pair of graphs can be mixed directly to generate an augmented graph. We conduct systematic experiments to show that S-Mixup can improve the performance and generalization of graph neural networks (GNNs) on various graph classification tasks. In addition, we show that S-Mixup can increase the robustness of GNNs against noisy labels. Our code is publicly available as part of the DIG package (https://github.com/divelab/DIG).
****

Deep Graph Representation Learning and Optimization for Influence Maximization
Chen Ling, Junji Jiang, Junxiang Wang, My T. Thai, Renhao Xue, James Song, Meikang Qiu, Liang Zhao
Influence maximization (IM) is formulated as selecting a set of initial users from a social network to maximize the expected number of influenced users. Researchers have made great progresses to design various traditional methods, yet both theoretical design and performance gain are close to their limits. In the past few years, learning-based IM methods have emerged to achieve stronger generalization ability to unknown graphs than traditional ones. However, the development of learning-based IM methods is still limited by fundamental obstacles, including 1) the difficulty of effectively solving the objective function; 2) the difficulty of characterizing the diversified and underlying diffusion patterns; and 3) the difficulty of adapting the solution under various node-centrality-constrained IM variants. To cope with the above challenges, we design a novel framework DeepIM to generatively characterize the latent representation of seed sets, and we propose to learn the diversified information diffusion pattern in a data-driven and end-to-end manner. Finally, we design a novel objective function to infer optimal seed sets under flexible node-centrality-based budget constraints. Extensive analyses are conducted over both synthetic and real-world datasets to demonstrate the overall performance of DeepIM.
****

Emergent Agentic Transformer from Chain of Hindsight Experience
Hao Liu, Pieter Abbeel
Large transformer models powered by diverse data and model scale have dominated natural language modeling and computer vision and pushed the frontier of multiple AI areas. In reinforcement learning (RL), despite many efforts into transformer-based policies, a key limitation, however, is that current transformer-based policies cannot learn by directly combining information from multiple sub-optimal trials. In this work, we address this issue using recently proposed chain of hindsight to relabel experience, where we train a transformer on a sequence of trajectory experience ascending sorted according to their total rewards. Our method consists of relabelling target return of each trajectory to the maximum total reward among in sequence of trajectories and training an autoregressive model to predict actions conditioning on past states, actions, rewards, target returns, and task completion tokens, the resulting model, Agentic Transformer (AT), can learn to improve upon itself both at training and test time. As we show on D4RL and ExoRL benchmarks, to the best our knowledge, this is the first time that a simple transformer-based model performs competitively with both temporal-difference and imitation-learning-based approaches, even from sub-optimal data. Our Agentic Transformer also shows a promising scaling trend that bigger models consistently improve results.
****

Shapley Based Residual Decomposition for Instance Analysis
Tommy Liu, Amanda S Barnard

In this paper, we introduce the idea of decomposing the residuals of regression with respect to the data instances instead of features. This allows us to determine the effects of each individual instance on the model and each other, and in doing so makes for a model-agnostic method of identifying instances of interest. In doing so, we can also determine the appropriateness of the model and data in the wider context of a given study. The paper focuses on the possible applications that such a framework brings to the relatively unexplored field of instance analysis in the context of Explainable AI tasks.
****
Learning Representations without Compositional Assumptions
Tennison Liu, Jeroen Berrevoets, Zhaozhi Qian, Mihaela Van Der Schaar
This paper addresses unsupervised representation learning on tabular data containing multiple views generated by distinct sources of measurement. Traditional methods, which tackle this problem using the multi-view framework, are constrained by predefined assumptions that assume feature sets share the same information and representations should learn globally shared factors. However, this assumption is not always valid for real-world tabular datasets with complex dependencies between feature sets, resulting in localized information that is harder to learn. To overcome this limitation, we propose a data-driven approach that learns feature set dependencies by representing feature sets as graph nodes and their relationships as learnable edges. Furthermore, we introduce $\texttt{LEGATO}$, a novel hierarchical graph autoencoder that learns a smaller, latent graph to aggregate information from multiple views dynamically. This approach results in latent graph components that specialize in capturing localized information from different regions of the input, leading to superior downstream performance.
****
Byzantine-Robust Learning on Heterogeneous Data via Gradient Splitting
Yuchen Liu, Chen Chen, Lingjuan Lyu, Fangzhao Wu, Sai Wu, Gang Chen
Federated learning has exhibited vulnerabilities to Byzantine attacks, where the Byzantine attackers can send arbitrary gradients to a central server to destroy the convergence and performance of the global model. A wealth of robust AGgregation Rules (AGRs) have been proposed to defend against Byzantine attacks. However, Byzantine clients can still circumvent robust AGRs when data is non-Identically and Independently Distributed (non-IID). In this paper, we first reveal the root causes of performance degradation of current robust AGRs in non-IID settings: the curse of dimensionality and gradient heterogeneity. In order to address this issue, we propose GAS, a GrAdient Splitting approach that can successfully adapt existing robust AGRs to non-IID settings. We also provide a detailed convergence analysis when the existing robust AGRs are combined with GAS. Experiments on various real-world datasets verify the efficacy of our proposed GAS. The implementation code is provided in https://github.com/YuchenLiu-a/byzantine-gas.
****
Towards Constituting Mathematical Structures for Learning to Optimize
Jialin Liu, Xiaohan Chen, Zhangyang Wang, Wotao Yin, Hanqin Cai
Learning to Optimize (L2O), a technique that utilizes machine learning to learn an optimization algorithm automatically from data, has gained arising attention in recent years. A generic L2O approach parameterizes the iterative update rule and learns the update direction as a black-box network. While the generic approach is widely applicable, the learned model can overfit and may not generalize well to out-of-distribution test sets. In this paper, we derive the basic mathematical conditions that successful update rules commonly satisfy. Consequently, we propose a novel L2O model with a mathematics-inspired structure that is broadly applicable and generalized well to out-of-distribution problems. Numerical simulations validate our theoretical findings and demonstrate the superior empirical performance of the proposed L2O model.
****
AudioLDM: Text-to-Audio Generation with Latent Diffusion Models
Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, Mark D Plumbley
Text-to-audio (TTA) systems have recently gained attention for their ability to

synthesize general audio based on text descriptions. However, previous studies in TTA have limited generation quality with high computational costs. In this study, we propose AudioLDM, a TTA system that is built on a latent space to learn continuous audio representations from contrastive language-audio pretraining (CLAP) embeddings. The pretrained CLAP models enable us to train LDMs with audio embeddings while providing text embeddings as the condition during sampling. By learning the latent representations of audio signals without modelling the cross-modal relationship, AudioLDM improves both generation quality and computational efficiency. Trained on AudioCaps with a single GPU, AudioLDM achieves state-of-the-art TTA performance compared to other open-sourced systems, measured by both objective and subjective metrics. AudioLDM is also the first TTA system that enables various text-guided audio manipulations (e.g., style transfer) in a zero-shot fashion. Our implementation and demos are available at https://audioldm.github.io.

****

Identifiability of Label Noise Transition Matrix

Yang Liu, Hao Cheng, Kun Zhang

The noise transition matrix plays a central role in the problem of learning with noisy labels. Among many other reasons, a large number of existing solutions rely on the knowledge of it. Identifying and estimating the transition matrix without ground truth labels is a critical and challenging task. When label noise transition depends on each instance, the problem of identifying the instance-dependent noise transition matrix becomes substantially more challenging. Despite recently proposed solutions for learning from instance-dependent noisy labels, the literature lacks a unified understanding of when such a problem remains identifiable. The goal of this paper is to characterize the identifiability of the label noise transition matrix. Building on Kruskal’s identifiability results, we are able to show the necessity of multiple noisy labels in identifying the noise transition matrix at the instance level. We further instantiate the results to explain the successes of the state-of-the-art solutions and how additional assumptions alleviated the requirement of multiple noisy labels. Our result reveals that disentangled features improve identification. This discovery led us to an approach that improves the estimation of the transition matrix using properly disentangled features. Code is available at https://github.com/UCSC-REAL/Identifiability.

****

A Group Symmetric Stochastic Differential Equation Model for Molecule Multi-modal Pretraining

Shengchao Liu, Weitao Du, Zhi-Ming Ma, Hongyu Guo, Jian Tang

Molecule pretraining has quickly become the go-to schema to boost the performance of AI-based drug discovery. Naturally, molecules can be represented as 2D topological graphs or 3D geometric point clouds. Although most existing pertaining methods focus on merely the single modality, recent research has shown that maximizing the mutual information (MI) between such two modalities enhances the molecule representation ability. Meanwhile, existing molecule multi-modal pretraining approaches approximate MI based on the representation space encoded from the topology and geometry, thus resulting in the loss of critical structural information of molecules. To address this issue, we propose MoleculeSDE. MoleculeSDE leverages group symmetric (e.g., SE(3)-equivariant and reflection-antisymmetric) stochastic differential equation models to generate the 3D geometries from 2D topologies, and vice versa, directly in the input space. It not only obtains tighter MI bound but also enables prosperous downstream tasks than the previous work. By comparing with 17 pretraining baselines, we empirically verify that MoleculeSDE can learn an expressive representation with state-of-the-art performance on 26 out of 32 downstream tasks.

****

Using Perturbation to Improve Goodness-of-Fit Tests based on Kernelized Stein Discrepancy

Xing Liu, Andrew B. Duncan, Axel Gandy

Kernelized Stein discrepancy (KSD) is a score-based discrepancy widely used in goodness-of-fit tests. It can be applied even when the target distribution has an

unknown normalising factor, such as in Bayesian analysis. We show theoretically and empirically that the KSD test can suffer from low power when the target and the alternative distributions have the same well-separated modes but differ in mixing proportions. We propose to perturb the observed sample via Markov transition kernels, with respect to which the target distribution is invariant. This allows us to then employ the KSD test on the perturbed sample. We provide numerical evidence that with suitably chosen transition kernels the proposed approach can lead to substantially higher power than the KSD test.

****

Cones: Concept Neurons in Diffusion Models for Customized Generation

Zhiheng Liu, Ruili Feng, Kai Zhu, Yifei Zhang, Kecheng Zheng, Yu Liu, Deli Zhao, Jingren Zhou, Yang Cao

Human brains respond to semantic features of presented stimuli with different neurons. This raises the question of whether deep neural networks admit a similar behavior pattern. To investigate this phenomenon, this paper identifies a small cluster of neurons associated with a specific subject in a diffusion model. We call those neurons the concept neurons. They can be identified by statistics of network gradients to a stimulation connected with the given subject. The concept neurons demonstrate magnetic properties in interpreting and manipulating generation results. Shutting them can directly yield the related subject contextualized in different scenes. Concatenating multiple clusters of concept neurons can vividly generate all related concepts in a single image. Our method attains impressive performance for multi-subject customization, even four or more subjects. For large-scale applications, the concept neurons are environmentally friendly as we only need to store a sparse cluster of int index instead of dense float32 parameter values, reducing storage consumption by 90% compared with previous customized generation methods. Extensive qualitative and quantitative studies on diverse scenarios show the superiority of our method in interpreting and manipulating diffusion models.

****

Opponent-Limited Online Search for Imperfect Information Games

Weiming Liu, Haobo Fu, Qiang Fu, Yang Wei

In recent years, online search has been playing an increasingly important role in imperfect information games (IIGs). Previous online search is known as common-knowledge subgame solving, which has to consider all the states in a common-knowledge closure. This is only computationally tolerable for medium size games, such as poker. To handle larger games, order-1 Knowledge-Limited Subgame Solving (1-KLSS) only considers the states in a knowledge-limited closure, which results in a much smaller subgame. However, 1-KLSS is unsafe. In this paper, we first extend 1-KLSS to Safe-1-KLSS and prove its safeness. To make Safe-1-KLSS applicable to even larger games, we propose Opponent-Limited Subgame Solving (OLSS) to limit how the opponent reaches a subgame and how it acts in the subgame. Limiting the opponent’s strategy dramatically reduces the subgame size and improves the efficiency of subgame solving while still preserving some safety in the limit. Experiments in medium size poker show that Safe-1-KLSS and OLSS are orders of magnitude faster than previous common-knowledge subgame solving. Also, OLSS significantly improves the online performance in a two-player Mahjong game, whose game size prohibits the use of previous common-knowledge subgame-solving methods.

****

Towards Robust and Safe Reinforcement Learning with Benign Off-policy Data

Zuxin Liu, Zijian Guo, Zhepeng Cen, Huan Zhang, Yihang Yao, Hanjiang Hu, Ding Zhao

Previous work demonstrates that the optimal safe reinforcement learning policy in a noise-free environment is vulnerable and could be unsafe under observational attacks. While adversarial training effectively improves robustness and safety, collecting samples by attacking the behavior agent online could be expensive or prohibitively dangerous in many applications. We propose the robuSt vAriational ofF-policy lEaRning (SAFER) approach, which only requires benign training data without attacking the agent. SAFER obtains an optimal non-parametric variational policy distribution via convex optimization and then uses it to improve the par

ameterized policy robustly via supervised learning. The two-stage policy optimiz ation facilitates robust training, and extensive experiments on multiple robot p latforms show the efficiency of SAFER in learning a robust and safe policy: achi eving the same reward with much fewer constraint violations during training than on-policy baselines.
****
Constrained Decision Transformer for Offline Safe Reinforcement Learning
Zuxin Liu, Zijian Guo, Yihang Yao, Zhepeng Cen, Wenhao Yu, Tingnan Zhang, Ding Z hao
Safe reinforcement learning (RL) trains a constraint satisfaction policy by inte racting with the environment. We aim to tackle a more challenging problem: learn ing a safe policy from an offline dataset. We study the offline safe RL problem from a novel multi-objective optimization perspective and propose the ϵ -reducible concept to characterize problem difficulties. The inherent trade-offs between safety and task performance inspire us to propose the constrained decis ion transformer (CDT) approach, which can dynamically adjust the trade-offs duri ng deployment. Extensive experiments show the advantages of the proposed method in learning an adaptive, safe, robust, and high-reward policy. CDT outperforms i ts variants and strong offline safe RL baselines by a large margin with the same hyperparameters across all tasks, while keeping the zero-shot adaptation capabi lity to different constraint thresholds, making our approach more suitable for r eal-world RL under constraints.
****
Understanding and Defending Patched-based Adversarial Attacks for Vision Transfo rmer
Liang Liu, Yanan Guo, Youtao Zhang, Jun Yang
Vision Transformer (ViT) is an attention-based model architecture that has demon strated superior performance on many computer vision tasks. However, its securit y properties, in particular, the robustness against adversarial attacks, are yet to be thoroughly studied. Recent works have shown that ViT is vulnerable to att ention-based adversarial patch attacks, which cover 1-3% area of the input image using adversarial patches and degrades the model accuracy to 0%. This work prov ides a generic study targeting the attention-based patch attack. First, we exper imentally observe that adversarial patches only activate in a few layers and bec ome lazy during attention updating. According to experiments, we study the theor y of how a small adversarial patch perturbates the whole model. Based on underst anding adversarial patch attacks, we propose a simple but efficient defense that correctly detects more than 95% of adversarial patches.
****
NUNO: A General Framework for Learning Parametric PDEs with Non-Uniform Data
Songming Liu, Zhongkai Hao, Chengyang Ying, Hang Su, Ze Cheng, Jun Zhu
The neural operator has emerged as a powerful tool in learning mappings between function spaces in PDEs. However, when faced with real-world physical data, whic h are often highly non-uniformly distributed, it is challenging to use mesh-base d techniques such as the FFT. To address this, we introduce the Non-Uniform Neur al Operator (NUNO), a comprehensive framework designed for efficient operator le arning with non-uniform data. Leveraging a K-D tree-based domain decomposition, we transform non-uniform data into uniform grids while effectively controlling i nterpolation error, thereby paralleling the speed and accuracy of learning from non-uniform data. We conduct extensive experiments on 2D elasticity, (2+1)D chan nel flow, and a 3D multi-physics heatsink, which, to our knowledge, marks a nove l exploration into 3D PDE problems with complex geometries. Our framework has re duced error rates by up to 60% and enhanced training speeds by 2x to 30x. The co de is now available at https://github.com/thu-ml/NUNO .
****
Hierarchical Programmatic Reinforcement Learning via Learning to Compose Program s
Guan-Ting Liu, En-Pei Hu, Pu-Jen Cheng, Hung-Yi Lee, Shao-Hua Sun
Aiming to produce reinforcement learning (RL) policies that are human-interpreta ble and can generalize better to novel scenarios, Trivedi et al. (2021) present

a method (LEAPS) that first learns a program embedding space to continuously parameterize diverse programs from a pre-generated program dataset, and then searches for a task-solving program in the learned program embedding space when given a task. Despite the encouraging results, the program policies that LEAPS can produce are limited by the distribution of the program dataset. Furthermore, during searching, LEAPS evaluates each candidate program solely based on its return, failing to precisely reward correct parts of programs and penalize incorrect parts. To address these issues, we propose to learn a meta-policy that composes a series of programs sampled from the learned program embedding space. By learning to compose programs, our proposed hierarchical programmatic reinforcement learning (HPRL) framework can produce program policies that describe out-of-distributionally complex behaviors and directly assign credits to programs that induce desired behaviors. The experimental results in the Karel domain show that our proposed framework outperforms baselines. The ablation studies confirm the limitations of LEAPS and justify our design choices.
****

Online Local Differential Private Quantile Inference via Self-normalization
Yi Liu, Qirui Hu, Lei Ding, Linglong Kong
Based on binary inquiries, we developed an algorithm to estimate population quantiles under Local Differential Privacy (LDP). By self-normalizing, our algorithm provides asymptotically normal estimation with valid inference, resulting in tight confidence intervals without the need for nuisance parameters to be estimated. Our proposed method can be conducted fully online, leading to high computational efficiency and minimal storage requirements with $\mathcal{O}(1)$ space. We also proved an optimality result by an elegant application of one central limit theorem of Gaussian Differential Privacy (GDP) when targeting the frequently encountered median estimation problem. With mathematical proof and extensive numerical testing, we demonstrate the validity of our algorithm both theoretically and experimentally.
****

GFlowOut: Dropout with Generative Flow Networks
Dianbo Liu, Moksh Jain, Bonaventure F. P. Dossou, Qianli Shen, Salem Lahlou, Anirudh Goyal, Nikolay Malkin, Chris Chinenye Emezue, Dinghuai Zhang, Nadhir Hassen, Xu Ji, Kenji Kawaguchi, Yoshua Bengio
Bayesian inference offers principled tools to tackle many critical problems with modern neural networks such as poor calibration and generalization, and data inefficiency. However, scaling Bayesian inference to large architectures is challenging and requires restrictive approximations. Monte Carlo Dropout has been widely used as a relatively cheap way to approximate inference and estimate uncertainty with deep neural networks. Traditionally, the dropout mask is sampled independently from a fixed distribution. Recent research shows that the dropout mask can be seen as a latent variable, which can be inferred with variational inference. These methods face two important challenges: (a) the posterior distribution over masks can be highly multi-modal which can be difficult to approximate with standard variational inference and (b) it is not trivial to fully utilize sample-dependent information and correlation among dropout masks to improve posterior estimation. In this work, we propose GFlowOut to address these issues. GFlowOut leverages the recently proposed probabilistic framework of Generative Flow Networks (GFlowNets) to learn the posterior distribution over dropout masks. We empirically demonstrate that GFlowOut results in predictive distributions that generalize better to out-of-distribution data and provide uncertainty estimates which lead to better performance in downstream tasks.
****

2D-Shapley: A Framework for Fragmented Data Valuation
Zhihong Liu, Hoang Anh Just, Xiangyu Chang, Xi Chen, Ruoxi Jia
Data valuation—quantifying the contribution of individual data sources to certain predictive behaviors of a model—is of great importance to enhancing the transparency of machine learning and designing incentive systems for data sharing. Existing work has focused on evaluating data sources with the shared feature or sample space. How to valuate fragmented data sources of which each only contains pa

rtial features and samples remains an open question. We start by presenting a me
thod to calculate the counterfactual of removing a fragment from the aggregated
data matrix. Based on the counterfactual calculation, we further propose 2D-Shap
ley, a theoretical framework for fragmented data valuation that uniquely satisfi
es some appealing axioms in the fragmented data context. 2D-Shapley empowers a r
ange of new use cases, such as selecting useful data fragments, providing interp
retation for sample-wise data values, and fine-grained data issue diagnosis.
****

Causal Structure Learning for Latent Intervened Non-stationary Data
Chenxi Liu, Kun Kuang
Causal structure learning can reveal the causal mechanism behind natural systems
. It is well studied that the multiple domain data consisting of observational a
nd interventional samples benefit causal identifiability. However, for non-stati
onary time series data, domain indexes are often unavailable, making it difficul
t to distinguish observational samples from interventional samples. To address t
hese issues, we propose a novel Latent Intervened Non-stationary learning (LIN)
method to make the domain indexes recovery process and the causal structure lear
ning process mutually promote each other. We characterize and justify a possible
 faithfulness condition to guarantee the identifiability of the proposed LIN met
hod. Extensive experiments on both synthetic and real-world datasets demonstrate
 that our method outperforms the baselines on causal structure learning for late
nt intervened non-stationary data.
****

Structural Re-weighting Improves Graph Domain Adaptation
Shikun Liu, Tianchun Li, Yongbin Feng, Nhan Tran, Han Zhao, Qiang Qiu, Pan Li
In many real-world applications, graph-structured data used for training and tes
ting have differences in distribution, such as in high energy physics (HEP) wher
e simulation data used for training may not match real experiments. Graph domain
 adaptation (GDA) is a method used to address these differences. However, curren
t GDA primarily works by aligning the distributions of node representations outp
ut by a single graph neural network encoder shared across the training and testi
ng domains, which may often yield sub-optimal solutions. This work examines diff
erent impacts of distribution shifts caused by either graph structure or node at
tributes and identifies a new type of shift, named conditional structure shift (
CSS), which current GDA approaches are provably sub-optimal to deal with. A nove
l approach, called structural reweighting (StruRW), is proposed to address this
issue and is tested on synthetic graphs, four benchmark datasets, and a new appl
ication in HEP. StruRW has shown significant performance improvement over the ba
selines in the settings with large graph structure shifts, and reasonable perfor
mance improvement when node attribute shift dominates.
****

Dink-Net: Neural Clustering on Large Graphs
Yue Liu, Ke Liang, Jun Xia, Sihang Zhou, Xihong Yang, Xinwang Liu, Stan Z. Li
Deep graph clustering, which aims to group the nodes of a graph into disjoint cl
usters with deep neural networks, has achieved promising progress in recent year
s. However, the existing methods fail to scale to the large graph with million n
odes. To solve this problem, a scalable deep graph clustering method (Dink-Net)
is proposed with the idea of dilation and shrink. Firstly, by discriminating nod
es, whether being corrupted by augmentations, representations are learned in a s
elf-supervised manner. Meanwhile, the cluster centers are initialized as learnab
le neural parameters. Subsequently, the clustering distribution is optimized by
minimizing the proposed cluster dilation loss and cluster shrink loss in an adve
rsarial manner. By these settings, we unify the two-step clustering, i.e., repre
sentation learning and clustering optimization, into an end-to-end framework, gu
iding the network to learn clustering-friendly features. Besides, Dink-Net scale
s well to large graphs since the designed loss functions adopt the mini-batch da
ta to optimize the clustering distribution even without performance drops. Both
experimental results and theoretical analyses demonstrate the superiority of our
 method. Compared to the runner-up, Dink-Net achieves $9.62%$ NMI improvement on
 the ogbn-papers100M dataset with 111 million nodes and 1.6 billion edges. The s

ource code is released: https://github.com/yueliu1999/Dink-Net. Besides, a colle
ction (papers, codes, and datasets) of deep graph clustering is shared on GitHub
 https://github.com/yueliu1999/Awesome-Deep-Graph-Clustering.
****

Oscillation-free Quantization for Low-bit Vision Transformers
Shih-Yang Liu, Zechun Liu, Kwang-Ting Cheng
Weight oscillation is a by-product of quantization-aware training, in which quan
tized weights frequently jump between two quantized levels, resulting in trainin
g instability and a sub-optimal final model. We discover that the learnable scal
ing factor, a widely-used $\textit{de facto}$ setting in quantization aggravates
 weight oscillation. In this work, we investigate the connection between learnab
le scaling factor and quantized weight oscillation using ViT, and we additionall
y find that the interdependence between quantized weights in $\textit{query}$ an
d $\textit{key}$ of a self-attention layer also makes ViT vulnerable to oscillat
ion. We propose three techniques correspondingly: statistical weight quantizatio
n ($\rm StatsQ$) to improve quantization robustness compared to the prevalent le
arnable-scale-based method; confidence-guided annealing ($\rm CGA$) that freezes
 the weights with $\textit{high confidence}$ and calms the oscillating weights;
and $\textit{query}$-$\textit{key}$ reparameterization ($\rm QKR$) to resolve th
e query-key intertwined oscillation and mitigate the resulting gradient misestim
ation. Extensive experiments demonstrate that our algorithms successfully abate
weight oscillation and consistently achieve substantial accuracy improvement on
ImageNet. Specifically, our 2-bit DeiT-T/DeiT-S surpass the previous state-of-th
e-art by 9.8% and 7.7%, respectively. The code is included in the supplementary
material and will be released.
****

Understanding the Distillation Process from Deep Generative Models to Tractable
Probabilistic Circuits
Xuejie Liu, Anji Liu, Guy Van Den Broeck, Yitao Liang
Probabilistic Circuits (PCs) are a general and unified computational framework f
or tractable probabilistic models that support efficient computation of various
inference tasks (e.g., computing marginal probabilities). Towards enabling such
reasoning capabilities in complex real-world tasks, Liu et al. (2022) propose to
 distill knowledge (through latent variable assignments) from less tractable but
 more expressive deep generative models. However, it is still unclear what facto
rs make this distillation work well. In this paper, we theoretically and empiric
ally discover that the performance of a PC can exceed that of its teacher model.
 Therefore, instead of performing distillation from the most expressive deep gen
erative model, we study what properties the teacher model and the PC should have
 in order to achieve good distillation performance. This leads to a generic algo
rithmic improvement as well as other data-type-specific ones over the existing l
atent variable distillation pipeline. Empirically, we outperform SoTA TPMs by a
large margin on challenging image modeling benchmarks. In particular, on ImageNe
t32, PCs achieve 4.06 bits-per-dimension, which is only 0.34 behind variational
diffusion models (Kingma et al., 2021).
****

Averaged Method of Multipliers for Bi-Level Optimization without Lower-Level Str
ong Convexity
Risheng Liu, Yaohua Liu, Wei Yao, Shangzhi Zeng, Jin Zhang
Gradient methods have become mainstream techniques for Bi-Level Optimization (BL
O) in learning fields. The validity of existing works heavily rely on either a r
estrictive Lower- Level Strong Convexity (LLSC) condition or on solving a series
 of approximation subproblems with high accuracy or both. In this work, by avera
ging the upper and lower level objectives, we propose a single loop Bi-level Ave
raged Method of Multipliers (sl-BAMM) for BLO that is simple yet efficient for l
arge-scale BLO and gets rid of the limited LLSC restriction. We further provide
non-asymptotic convergence analysis of sl-BAMM towards KKT stationary points, an
d the comparative advantage of our analysis lies in the absence of strong gradie
nt boundedness assumption, which is always required by others. Thus our theory s
afely captures a wider variety of applications in deep learning, especially wher

e the upper-level objective is quadratic w.r.t. the lower-level variable. Experi
mental results demonstrate the superiority of our method.
****

Graph Switching Dynamical Systems
Yongtuo Liu, Sara Magliacane, Miltiadis Kofinas, Efstratios Gavves
Dynamical systems with complex behaviours, e.g. immune system cells interacting
with a pathogen, are commonly modelled by splitting the behaviour in different r
egimes, or modes, each with simpler dynamics, and then learn the switching behav
iour from one mode to another. To achieve this, Switching Dynamical Systems (SDS
) are a powerful tool that automatically discovers these modes and mode-switchin
g behaviour from time series data. While effective, these methods focus on indep
endent objects, where the modes of one object are independent of the modes of th
e other objects. In this paper, we focus on the more general interacting object
setting for switching dynamical systems, where the per-object dynamics also depe
nd on an unknown and dynamically changing subset of other objects and their mode
s. To this end, we propose a novel graph-based approach for switching dynamical
systems, GRAph Switching dynamical Systems (GRASS), in which we use a dynamic gr
aph to characterize interactions between objects and learn both intra-object and
 inter-object mode-switching behaviour. For benchmarking, we create two new data
sets, a synthesized ODE-driven particles dataset and a real-world Salsa-couple d
ancing dataset. Experiments show that GRASS can consistently outperforms previou
s state-of-the-art methods. We will release code and data after acceptance.
****

High Probability Convergence of Stochastic Gradient Methods
Zijian Liu, Ta Duy Nguyen, Thien Hang Nguyen, Alina Ene, Huy Nguyen
In this work, we describe a generic approach to show convergence with high proba
bility for both stochastic convex and non-convex optimization with sub-Gaussian
noise. In previous works for convex optimization, either the convergence is only
 in expectation or the bound depends on the diameter of the domain. Instead, we
show high probability convergence with bounds depending on the initial distance
to the optimal solution. The algorithms use step sizes analogous to the standard
 settings and are universal to Lipschitz functions, smooth functions, and their
linear combinations. The method can be applied to the non-convex case. We demons
trate an $O((1+\sigma^{2}\log(1/\delta))/T+\sigma/\sqrt{T})$ convergence rate wh
en the number of iterations T is known and an $O((1+\sigma^{2}\log(T/\delta))/
\sqrt{T})$ convergence rate when T is unknown for SGD, where $1-\delta$ is the
 desired success probability. These bounds improve over existing bounds in the l
iterature. We also revisit AdaGrad-Norm (Ward et al., 2019) and show a new analy
sis to obtain a high probability bound that does not require the bounded gradien
t assumption made in previous works. The full version of our paper contains resu
lts for the standard per-coordinate AdaGrad.
****

OMS-DPM: Optimizing the Model Schedule for Diffusion Probabilistic Models
Enshu Liu, Xuefei Ning, Zinan Lin, Huazhong Yang, Yu Wang
Diffusion probabilistic models (DPMs) are a new class of generative models that
have achieved state-of-the-art generation quality in various domains. Despite th
e promise, one major drawback of DPMs is the slow generation speed due to the la
rge number of neural network evaluations required in the generation process. In
this paper, we reveal an overlooked dimension—model schedule—for optimizing the
trade-off between generation quality and speed. More specifically, we observe th
at small models, though having worse generation quality when used alone, could o
utperform large models in certain generation steps. Therefore, unlike the tradit
ional way of using a single model, using different models in different generatio
n steps in a carefully designed model schedule could potentially improve generat
ion quality and speed simultaneously. We design OMS-DPM, a predictor-based searc
h algorithm, to determine the optimal model schedule given an arbitrary generati
on time budget and a set of pre-trained models. We demonstrate that OMS-DPM can
find model schedules that improve generation quality and speed than prior state-
of-the-art methods across CIFAR-10, CelebA, ImageNet, and LSUN datasets. When ap
plied to the public checkpoints of the Stable Diffusion model, we are able to ac

celerate the sampling by 2x while maintaining the generation quality.
****
Lazy Agents: A New Perspective on Solving Sparse Reward Problem in Multi-agent Reinforcement Learning
Boyin Liu, Zhiqiang Pu, Yi Pan, Jianqiang Yi, Yanyan Liang, D. Zhang
Sparse reward remains a valuable and challenging problem in multi-agent reinforcement learning (MARL). This paper addresses this issue from a new perspective, i.e., lazy agents. We empirically illustrate how lazy agents damage learning from both exploration and exploitation. Then, we propose a novel MARL framework called Lazy Agents Avoidance through Influencing External States (LAIES). Firstly, we examine the causes and types of lazy agents in MARL using a causal graph of the interaction between agents and their environment. Then, we mathematically define the concept of fully lazy agents and teams by calculating the causal effect of their actions on external states using the do-calculus process. Based on definitions, we provide two intrinsic rewards to motivate agents, i.e., individual diligence intrinsic motivation (IDI) and collaborative diligence intrinsic motivation (CDI). IDI and CDI employ counterfactual reasoning based on the external states transition model (ESTM) we developed. Empirical results demonstrate that our proposed method achieves state-of-the-art performance on various tasks, including the sparse-reward version of StarCraft multi-agent challenge (SMAC) and Google Research Football (GRF). Our code is open-source and available at https://github.com/liuboyin/LAIES.
****
RSC: Accelerate Graph Neural Networks Training via Randomized Sparse Computations
Zirui Liu, Chen Shengyuan, Kaixiong Zhou, Daochen Zha, Xiao Huang, Xia Hu
Training graph neural networks (GNNs) is extremely time consuming because sparse graph-based operations are hard to be accelerated by community hardware. Prior art successfully reduces the computation cost of dense matrix based operations (e.g., convolution and linear) via sampling-based approximation. However, unlike dense matrices, sparse matrices are stored in the irregular data format such that each row/column may have different number of non-zero entries. Thus, compared to the dense counterpart, approximating sparse operations has two unique challenges (1) we cannot directly control the efficiency of approximated sparse operation since the computation is only executed on non-zero entries; (2) sampling sparse matrices is much more inefficient due to the irregular data format. To address the issues, our key idea is to control the accuracy-efficiency trade off by optimizing computation resource allocation layer-wisely and epoch-wisely. For the first challenge, we customize the computation resource to different sparse operations, while limit the total used resource below a certain budget. For the second challenge, we cache previous sampled sparse matrices to reduce the epoch-wise sampling overhead. Finally, we propose a switching mechanisms to improve the generalization of GNNs trained with approximated operations. To this end, we propose Randomized Sparse Computation. In practice, rsc can achieve up to 11.6X speedup for a single sparse operation and 1.6X end-to-end wall-clock time speedup with almost no accuracy drop.
****
Algorithms for bounding contribution for histogram estimation under user-level privacy
Yuhan Liu, Ananda Theertha Suresh, Wennan Zhu, Peter Kairouz, Marco Gruteser
We study the problem of histogram estimation under user-level differential privacy, where the goal is to preserve the privacy of all entries of any single user. We consider the heterogeneous scenario where the quantity of data can be different for each user. In this scenario, the amount of noise injected into the histogram to obtain differential privacy is proportional to the maximum user contribution, which can be amplified by few outliers. One approach to circumvent this would be to bound (or limit) the contribution of each user to the histogram. However, if users are limited to small contributions, a significant amount of data will be discarded. In this work, we propose algorithms to choose the best user contribution bound for histogram estimation under both bounded and unbounded domain

settings. When the size of the domain is bounded, we propose a user contributio n bounding strategy that almost achieves a two-approximation with respect to the best contribution bound in hindsight. For unbounded domain histogram estimation , we propose an algorithm that is logarithmic-approximation with respect to the best contribution bound in hindsight. This result holds without any distribution assumptions on the data. Experiments on both real and synthetic datasets verify our theoretical findings and demonstrate the effectiveness of our algorithms. W e also show that clipping bias introduced by bounding user contribution may be r educed under mild distribution assumptions, which can be of independent interest .

****

Simple Embodied Language Learning as a Byproduct of Meta-Reinforcement Learning

Evan Zheran Liu, Sahaana Suri, Tong Mu, Allan Zhou, Chelsea Finn

Whereas machine learning models typically learn language by directly training on language tasks (e.g., next-word prediction), language emerges in human children as a byproduct of solving non-language tasks (e.g., acquiring food). Motivated by this observation, we ask: can embodied reinforcement learning (RL) agents als o indirectly learn language from non-language tasks? Learning to associate langu age with its meaning requires a dynamic environment with varied language. Theref ore, we investigate this question in a multi-task environment with language that varies across the different tasks. Specifically, we design an office navigation environment, where the agent’s goal is to find a particular office, and office locations differ in different buildings (i.e., tasks). Each building includes a floor plan with a simple language description of the goal office’s location, whi ch can be visually read as an RGB image when visited. We find RL agents indeed a re able to indirectly learn language. Agents trained with current meta-RL algori thms successfully generalize to reading floor plans with held-out layouts and la nguage phrases, and quickly navigate to the correct office, despite receiving no direct language supervision.

****

Generating Private Synthetic Data with Genetic Algorithms

Terrance Liu, Jingwu Tang, Giuseppe Vietri, Steven Wu

We study the problem of efficiently generating differentially private synthetic data that approximate the statistical properties of an underlying sensitive data set. In recent years, there has been a growing line of work that approaches this problem using first-order optimization techniques. However, such techniques are restricted to optimizing differentiable objectives only, severely limiting the types of analyses that can be conducted. For example, first-order mechanisms hav e been primarily successful in approximating statistical queries only in the for m of marginals for discrete data domains. In some cases, one can circumvent such issues by relaxing the task’s objective to maintain differentiability. However, even when possible, these approaches impose a fundamental limitation in which m odifications to the minimization problem become additional sources of error. The refore, we propose Private-GSD, a private genetic algorithm based on zeroth-orde r optimization heuristics that do not require modifying the original objective; thus, it avoids the aforementioned limitations of first-order optimization. We d emonstrate empirically that on data with both discrete and real-valued attribute s, Private-GSD outperforms the state-of-the-art methods on non-differential quer ies while matching accuracy in approximating differentiable ones.

****

FusionRetro: Molecule Representation Fusion via In-Context Learning for Retrosyn thetic Planning

Songtao Liu, Zhengkai Tu, Minkai Xu, Zuobai Zhang, Lu Lin, Rex Ying, Jian Tang, Peilin Zhao, Dinghao Wu

Retrosynthetic planning aims to devise a complete multi-step synthetic route fro m starting materials to a target molecule. Current strategies use a decoupled ap proach of single-step retrosynthesis models and search algorithms, taking only t he product as the input to predict the reactants for each planning step and igno ring valuable context information along the synthetic route. In this work, we pr opose a novel framework that utilizes context information for improved retrosynt

hetic planning. We view synthetic routes as reaction graphs and propose to incorporate context through three principled steps: encode molecules into embeddings, aggregate information over routes, and readout to predict reactants. Our approach is the first attempt to utilize in-context learning for retrosynthesis prediction in retrosynthetic planning. The entire framework can be efficiently optimized in an end-to-end fashion and produce more practical and accurate predictions. Comprehensive experiments demonstrate that by fusing in the context information over routes, our model significantly improves the performance of retrosynthetic planning over baselines that are not context-aware, especially for long synthetic routes. Code is available at https://github.com/SongtaoLiu0823/FusionRetro.
****

I^{2}SB: Image-to-Image Schrödinger Bridge
Guan-Horng Liu, Arash Vahdat, De-An Huang, Evangelos Theodorou, Weili Nie, Anima Anandkumar
We propose Image-to-Image Schrödinger Bridge (I^{2}SB), a new class of conditional diffusion models that directly learn the nonlinear diffusion processes between two given distributions. These diffusion bridges are particularly useful for image restoration, as the degraded images are structurally informative priors for reconstructing the clean images. I^{2}SB belongs to a tractable class of Schrödinger bridge, the nonlinear extension to score-based models, whose marginal distributions can be computed analytically given boundary pairs. This results in a simulation-free framework for nonlinear diffusions, where the I^{2}SB training becomes scalable by adopting practical techniques used in standard diffusion models. We validate I^{2}SB in solving various image restoration tasks, including inpainting, super-resolution, deblurring, and JPEG restoration on ImageNet 256$\times$256 and show that I^{2}SB surpasses standard conditional diffusion models with more interpretable generative processes. Moreover, I^{2}SB matches the performance of inverse methods that additionally require the knowledge of the corruption operators. Our work opens up new algorithmic opportunities for developing efficient nonlinear diffusion models on a large scale. Project page and codes: https://i2sb.github.io/
****

What can online reinforcement learning with function approximation benefit from general coverage conditions?
Fanghui Liu, Luca Viano, Volkan Cevher
In online reinforcement learning (RL), instead of employing standard structural assumptions on Markov decision processes (MDPs), using a certain coverage condition (original from offline RL) is enough to ensure sample-efficient guarantees (Xie et al. 2023). In this work, we focus on this new direction by digging more possible and general coverage conditions, and study the potential and the utility of them in efficient online RL. We identify more concepts, including the L^p variant of concentrability, the density ratio realizability, and trade-off on the partial/rest coverage condition, that can be also beneficial to sample-efficient online RL, achieving improved regret bound. Furthermore, if exploratory offline data are used, under our coverage conditions, both statistically and computationally efficient guarantees can be achieved for online RL. Besides, even though the MDP structure is given, e.g., linear MDP, we elucidate that, good coverage conditions are still beneficial to obtain faster regret bound beyond $\widetilde{\mathcal{O}}(\sqrt{T})$ and even a logarithmic order regret. These results provide a good justification for the usage of general coverage conditions in efficient online RL.
****

TR0N: Translator Networks for 0-Shot Plug-and-Play Conditional Generation
Zhaoyan Liu, Noël Vouitsis, Satya Krishna Gorti, Jimmy Ba, Gabriel Loaiza-Ganem
We propose TR0N, a highly general framework to turn pre-trained unconditional generative models, such as GANs and VAEs, into conditional models. The conditioning can be highly arbitrary, and requires only a pre-trained auxiliary model. For example, we show how to turn unconditional models into class-conditional ones with the help of a classifier, and also into text-to-image models by leveraging CLIP. TR0N learns a lightweight stochastic mapping which "translates’" between the

space of conditions and the latent space of the generative model, in such a way that the generated latent corresponds to a data sample satisfying the desired condition. The translated latent samples are then further improved upon through Langevin dynamics, enabling us to obtain higher-quality data samples. TR0N requires no training data nor fine-tuning, yet can achieve a zero-shot FID of 10.9 on MS-COCO, outperforming competing alternatives not only on this metric, but also in sampling speed – all while retaining a much higher level of generality. Our code is available at https://github.com/layer6ai-labs/tr0n.

****

Global Optimization with Parametric Function Approximation

Chong Liu, Yu-Xiang Wang

We consider the problem of global optimization with noisy zeroth order oracles — a well-motivated problem useful for various applications ranging from hyper-parameter tuning for deep learning to new material design. Existing work relies on Gaussian processes or other non-parametric family, which suffers from the curse of dimensionality. In this paper, we propose a new algorithm GO-UCB that leverages a parametric family of functions (e.g., neural networks) instead. Under a realizable assumption and a few other mild geometric conditions, we show that GO-UCB achieves a cumulative regret of $\tilde{O}(\sqrt{T})$ where T is the time horizon. At the core of GO-UCB is a carefully designed uncertainty set over parameters based on gradients that allows optimistic exploration. Synthetic and real-world experiments illustrate GO-UCB works better than popular Bayesian optimization approaches, even if the model is misspecified.

****

Deja Vu: Contextual Sparsity for Efficient LLMs at Inference Time

Zichang Liu, Jue Wang, Tri Dao, Tianyi Zhou, Binhang Yuan, Zhao Song, Anshumali Shrivastava, Ce Zhang, Yuandong Tian, Christopher Re, Beidi Chen

Large language models (LLMs) with hundreds of billions of parameters have sparked a new wave of exciting AI applications. However, they are computationally expensive at inference time. Sparsity is a natural approach to reduce this cost, but existing methods either require costly retraining, have to forgo LLM’s in-context learning ability, or do not yield wall-clock time speedup on modern hardware. We hypothesize that contextual sparsity, which are small, input-dependent sets of attention heads and MLP parameters that yield approximately the same output as the dense model for a given input, can address these issues. We show that contextual sparsity exists, that it can be accurately predicted, and that we can exploit it to speed up LLM inference in wall-clock time without compromising LLM’s quality or in-context learning ability. Based on these insights, we propose Deja Vu, a system that uses a low-cost algorithm to predict contextual sparsity on the fly given inputs to each layer, along with an asynchronous and hardware-aware implementation that speeds up LLM inference. We validate that DejaVu can reduce the inference latency of OPT-175B by over 2$\times$ compared to the state-of-the-art FasterTransformer, and over 6$\times$ compared to the widely used Hugging Face implementation, without compromising model quality. The code is available at https://github.com/FMInference/DejaVu.

****

Trapdoor Normalization with Irreversible Ownership Verification

Hanwen Liu, Zhenyu Weng, Yuesheng Zhu, Yadong Mu

This paper introduces a deep model watermark with an irreversible ownership verification scheme: Trapdoor Normalization (TdN), inspired by the trapdoor function in traditional cryptography. To protect intellectual property within deep models, the proposed method is able to embed ownership information into normalization layers during training. We argue and empirically validate that relevant methods are vulnerable to ambiguity attacks, where the forged watermarks can cast ambiguity over the ownership verification. The primary trait that distinguishes this work from previous ones, is its design of a bidirectional connection between watermarks and deep models. Thereby, TdN enables an irreversible ownership verification scheme that is difficult for the adversary to compromise. In this way, the proposed TdN can effectively defeat ambiguity attacks. Extensive experiments demonstrate that the proposed method is not only superior to previous state-of-the-

art methods in robustness, but also has better efficiency.
****
Same Pre-training Loss, Better Downstream: Implicit Bias Matters for Language Models
Hong Liu, Sang Michael Xie, Zhiyuan Li, Tengyu Ma
Language modeling on large-scale datasets improves performance of various downstream tasks. The validation pre-training loss is often used as the evaluation metric for language models since the pre-training loss tends to be well-correlated with downstream performance (which is itself hard to evaluate comprehensively). Contrary to the conventional wisdom, this paper shows that 1) pre-training loss cannot fully explain downstream performance and 2) flatness of the model is well-correlated with downstream performance where pre-training loss is not. We identify three ways to produce models with the same pre-training loss but different downstream performance: continue pre-training after convergence, increasing the model size, and changing the pre-training algorithms. These experiments demonstrate the existence of implicit bias of pre-training algorithms—among models with the same minimal pre-training loss, they implicitly prefer more transferable ones. Toward understanding this implicit bias, we prove that SGD with standard mini-batch noise implicitly prefers flatter minima of pre-training loss in language models, and empirically observe a strong correlation between flatness (measured by the trace of Hessian) and downstream performance among models with the same pre-training loss. We also prove in a synthetic language setting that among models with the minimal pre-training loss, the flattest model transfers to downstream tasks.
****
Taxonomy-Structured Domain Adaptation
Tianyi Liu, Zihao Xu, Hao He, Guang-Yuan Hao, Guang-He Lee, Hao Wang
Domain adaptation aims to mitigate distribution shifts among different domains. However, traditional formulations are mostly limited to categorical domains, greatly simplifying nuanced domain relationships in the real world. In this work, we tackle a generalization with taxonomy-structured domains, which formalizes domains with nested, hierarchical similarity structures such as animal species and product catalogs. We build on the classic adversarial framework and introduce a novel taxonomist, which competes with the adversarial discriminator to preserve the taxonomy information. The equilibrium recovers the classic adversarial domain adaptation’s solution if given a non-informative domain taxonomy (e.g., a flat taxonomy where all leaf nodes connect to the root node) while yielding non-trivial results with other taxonomies. Empirically, our method achieves state-of-the-art performance on both synthetic and real-world datasets with successful adaptation.
****
Dropout Reduces Underfitting
Zhuang Liu, Zhiqiu Xu, Joseph Jin, Zhiqiang Shen, Trevor Darrell
Introduced by Hinton et al. in 2012, dropout has stood the test of time as a regularizer for preventing overfitting in neural networks. In this study, we demonstrate that dropout can also mitigate underfitting when used at the start of training. During the early phase, we find dropout reduces the directional variance of gradients across mini-batches and helps align the mini-batch gradients with the entire dataset’s gradient. This helps counteract the stochasticity of SGD and limit the influence of individual batches on model training. Our findings lead us to a solution for improving performance in underfitting models - early dropout: dropout is applied only during the initial phases of training, and turned off afterwards. Models equipped with early dropout achieve lower final training loss compared to their counterparts without dropout. Additionally, we explore a symmetric technique for regularizing overfitting models - late dropout, where dropout is not used in the early iterations and is only activated later in training. Experiments on ImageNet and various vision tasks demonstrate that our methods consistently improve generalization accuracy. Our results encourage more research on understanding regularization in deep learning and our methods can be useful tools for future neural network training, especially in the era of large data. Cod

e is available at https://github.com/facebookresearch/dropout.
****
Revisiting Pseudo-Label for Single-Positive Multi-Label Learning
Biao Liu, Ning Xu, Jiaqi Lv, Xin Geng
To deal with the challenge of high cost of annotating all relevant labels for each example in multi-label learning, single-positive multi-label learning (SPMLL) has been studied in recent years, where each example is annotated with only one positive label. By adopting pseudo-label generation, i.e., assigning pseudo-label to each example by various strategies, existing methods have empirically validated that SPMLL would significantly reduce the amount of supervision with a tolerable damage in classification performance. However, there is no existing method that can provide a theoretical guarantee for learning from pseudo-label on SPMLL. In this paper, the conditions of the effectiveness of learning from pseudo-label for SPMLL are shown and the learnability of pseudo-label-based methods is proven. Furthermore, based on the theoretical guarantee of pseudo-label for SPMLL, we propose a novel SPMLL method named MIME, i.e., Mutual label enhancement for sIngle-positive Multi-label lEarning and prove that the generated pseudo-label by MIME approximately converges to the fully-supervised case. Experiments on four image datasets and five MLL datasets show the effectiveness of our methods over several existing SPMLL approaches.
****
Retrosynthetic Planning with Dual Value Networks
Guoqing Liu, Di Xue, Shufang Xie, Yingce Xia, Austin Tripp, Krzysztof Maziarz, Marwin Segler, Tao Qin, Zongzhang Zhang, Tie-Yan Liu
Retrosynthesis, which aims to find a route to synthesize a target molecule from commercially available starting materials, is a critical task in drug discovery and materials design. Recently, the combination of ML-based single-step reaction predictors with multi-step planners has led to promising results. However, the single-step predictors are mostly trained offline to optimize the single-step accuracy, without considering complete routes. Here, we leverage reinforcement learning (RL) to improve the single-step predictor, by using a tree-shaped MDP to optimize complete routes. Specifically, we propose a novel online training algorithm, called Planning with Dual Value Networks (PDVN), which alternates between the planning phase and updating phase. In PDVN, we construct two separate value networks to predict the synthesizability and cost of molecules, respectively. To maintain the single-step accuracy, we design a two-branch network structure for the single-step predictor. On the widely-used USPTO dataset, our PDVN algorithm improves the search success rate of existing multi-step planners (e.g., increasing the success rate from 85.79% to 98.95% for Retro*, and reducing the number of model calls by half while solving 99.47% molecules for RetroGraph). Additionally, PDVN helps find shorter synthesis routes (e.g., reducing the average route length from 5.76 to 4.83 for Retro*, and from 5.63 to 4.78 for RetroGraph).
****
Online Nonstochastic Control with Adversarial and Static Constraints
Xin Liu, Zixian Yang, Lei Ying
This paper studies online nonstochastic control problems with adversarial and static constraints. We propose online nonstochastic control algorithms that achieve both sublinear regret and sublinear adversarial constraint violation while keeping static constraint violation minimal against the optimal constrained linear control policy in hindsight. To establish the results, we introduce an online convex optimization with memory framework under adversarial and static constraints, which serves as a subroutine for the constrained online nonstochastic control algorithms. This subroutine also achieves the state-of-the-art regret and constraint violation bounds for constrained online convex optimization problems, which is of independent interest. Our experiments demonstrate the proposed control algorithms are adaptive to adversarial constraints and achieve smaller cumulative costs and violations. Moreover, our algorithms are less conservative and achieve significantly smaller cumulative costs than the state-of-the-art algorithm.
****

Optimization for Amortized Inverse Problems
Tianci Liu, Tong Yang, Quan Zhang, Qi Lei
Incorporating a deep generative model as the prior distribution in inverse problems has established substantial success in reconstructing images from corrupted observations. Notwithstanding, the existing optimization approaches use gradient descent largely without adapting to the non-convex nature of the problem and can be sensitive to initial values, impeding further performance improvement. In this paper, we propose an efficient amortized optimization scheme for inverse problems with a deep generative prior. Specifically, the optimization task with high degrees of difficulty is decomposed into optimizing a sequence of much easier ones. We provide a theoretical guarantee of the proposed algorithm and empirically validate it on different inverse problems. As a result, our approach outperforms baseline methods qualitatively and quantitatively by a large margin.
****
Active Policy Improvement from Multiple Black-box Oracles
Xuefeng Liu, Takuma Yoneda, Chaoqi Wang, Matthew Walter, Yuxin Chen
Reinforcement learning (RL) has made significant strides in various complex domains. However, identifying an effective policy via RL often necessitates extensive exploration. Imitation learning aims to mitigate this issue by using expert demonstrations to guide exploration. In real-world scenarios, one often has access to multiple suboptimal black-box experts, rather than a single optimal oracle. These experts do not universally outperform each other across all states, presenting a challenge in actively deciding which oracle to use and in which state. We introduce MAPS and MAPS-SE, a class of policy improvement algorithms that perform imitation learning from multiple suboptimal oracles. In particular, MAPS actively selects which of the oracles to imitate and improve their value function estimates, and MAPS-SE additionally leverages an active state exploration criterion to determine which states one should explore. We provide a comprehensive theoretical analysis and demonstrate that MAPS and MAPS-SE enjoy sample efficiency advantage over the state-of-the-art policy improvement algorithms. Empirical results show that MAPS-SE significantly accelerates policy optimization via state-wise imitation learning from multiple oracles across a broad spectrum of control tasks in the DeepMind Control Suite.
****
Gradient-based Wang-Landau Algorithm: A Novel Sampler for Output Distribution of Neural Networks over the Input Space
Weitang Liu, Yi-Zhuang You, Ying Wai Li, Jingbo Shang
The output distribution of a neural network (NN) over the entire input space captures the complete input-output mapping relationship, offering in- sights toward a more comprehensive NN under- standing. Exhaustive enumeration or traditional Monte Carlo methods for the entire input space can exhibit impractical sampling time, especially for high-dimensional inputs. To make such difficult sampling computationally feasible, in this paper, we propose a novel Gradient-based Wang-Landau (GWL) sampler. We first draw the connection between the output distribution of a NN and the density of states (DOS) of a physical system. Then, we renovate the classic sampler for the DOS problem, Wang-Landau algorithm, by re-placing its random proposals with gradient-based Monte Carlo proposals. This way, our GWL sampler investigates the under-explored subsets of the input space much more efficiently. Extensive experiments have verified the accuracy of the output distribution generated by GWL and also showcased several interesting findings - for example, in a binary image classification task, both CNN and ResNet mapped the majority of human unrecognizable images to very negative logit values.
****
VectorMapNet: End-to-end Vectorized HD Map Learning
Yicheng Liu, Tianyuan Yuan, Yue Wang, Yilun Wang, Hang Zhao
Autonomous driving systems require High-Definition (HD) semantic maps to navigate around urban roads. Existing solutions approach the semantic mapping problem by offline manual annotation, which suffers from serious scalability issues. Recent learning-based methods produce dense rasterized segmentation predictions to construct maps. However, these predictions do not include instance information of

individual map elements and require heuristic post-processing to obtain vectorized maps. To tackle these challenges, we introduce an end-to-end vectorized HD map learning pipeline, termed VectorMapNet. VectorMapNet takes onboard sensor observations and predicts a sparse set of polylines in the bird’s-eye view. This pipeline can explicitly model the spatial relation between map elements and generate vectorized maps that are friendly to downstream autonomous driving tasks. Extensive experiments show that VectorMapNet achieve strong map learning performance on both nuScenes and Argoverse2 dataset, surpassing previous state-of-the-art methods by 14.2 mAP and 14.6mAP. Qualitatively, VectorMapNet is capable of generating comprehensive maps and capturing fine-grained details of road geometry. To the best of our knowledge, VectorMapNet is the first work designed towards end-to-end vectorized map learning from onboard observations.

****

Partially Observable Multi-agent RL with (Quasi-)Efficiency: The Blessing of Information Sharing

Xiangyu Liu, Kaiqing Zhang

We study provable multi-agent reinforcement learning (MARL) in the general framework of partially observable stochastic games (POSGs). To circumvent the known hardness results and the use of computationally intractable oracles, we propose to leverage the potential information-sharing among agents, a standard practice in empirical MARL and a common model for multi-agent control systems with communications. We first establish several computation complexity results to justify the necessity of information-sharing, as well as the observability assumption that has enabled quasi-efficient single-agent RL with partial observations, for computational efficiency in solving POSGs. We then propose to further approximate the shared common information to construct an approximate model of the POSG, in which planning an approximate equilibrium (in terms of solving the original POSG) can be quasi-efficient, i.e., of quasi-polynomial-time, under the aforementioned assumptions. Furthermore, we develop a partially observable MARL algorithm that is both statistically and computationally quasi-efficient. We hope our study can open up the possibilities of leveraging and even designing different information structures, for developing both sample- and computation-efficient partially observable MARL.

****

Prometheus: Taming Sample and Communication Complexities in Constrained Decentralized Stochastic Bilevel Learning

Zhuqing Liu, Xin Zhang, Prashant Khanduri, Songtao Lu, Jia Liu

In recent years, decentralized bilevel optimization has gained significant attention thanks to its versatility in modeling a wide range of multi-agent learning problems, such as multi-agent reinforcement learning and multi-agent meta-learning. However, one unexplored and fundamental problem in this area is how to solve decentralized stochastic bilevel optimization problems with domain constraints while achieving low sample and communication complexities. This problem often arises from multi-agent learning problems with safety constraints. As shown in this paper, constrained decentralized bilevel optimization is far more challenging than its unconstrained counterpart due to the complex coupling structure, which necessitates new algorithm design and analysis techniques. Toward this end, we investigate a class of constrained decentralized bilevel optimization problems, where multiple agents collectively solve a nonconvex-strongly-convex bilevel problem with constraints in the upper-level variables. We propose an algorithm called Prometheus (proximal tracked stochastic recursive estimator) that achieves the first $\mathcal{O}(\epsilon^{-1})$ results in both sample and communication complexities for constrained decentralized bilevel optimization, where $\epsilon>0$ is a desired stationarity error. Collectively, the results in this work contribute to a theoretical foundation for low sample- and communication-complexity constrained decentralized bilevel learning.

****

D2Match: Leveraging Deep Learning and Degeneracy for Subgraph Matching

Xuanzhou Liu, Lin Zhang, Jiaqi Sun, Yujiu Yang, Haiqin Yang

Subgraph matching is a fundamental building block for graph-based applications a

nd is challenging due to its high-order combinatorial nature. Existing studies u
sually tackle it by combinatorial optimization or learning-based methods. Howeve
r, they suffer from exponential computational costs or searching the matching wi
thout theoretical guarantees. In this paper, we develop D^2Match by leveraging
 the efficiency of Deep learning and Degeneracy for subgraph matching. More spec
ifically, we first prove that subgraph matching can degenerate to subtree matchi
ng, and subsequently is equivalent to finding a perfect matching on a bipartite
graph. We can then yield an implementation of linear time complexity by the buil
t-in tree-structured aggregation mechanism on graph neural networks. Moreover, c
ircle structures and node attributes can be easily incorporated in D^2Match to
 boost the matching performance. Finally, we conduct extensive experiments to sh
ow the superior performance of our D^2Match and confirm that our D^2Match in
deed exploits the subtrees and differs from existing GNNs-based subgraph matchin
g methods that depend on memorizing the data distribution divergence.
****

Image Shortcut Squeezing: Countering Perturbative Availability Poisons with Comp
ression

Zhuoran Liu, Zhengyu Zhao, Martha Larson

Perturbative availability poisoning (PAP) adds small changes to images to preven
t their use for model training. Current research adopts the belief that practica
l and effective approaches to countering such poisons do not exist. In this pape
r, we argue that it is time to abandon this belief. We present extensive experim
ents showing that 12 state-of-the-art PAP methods are vulnerable to Image Shortc
ut Squeezing (ISS), which is based on simple compression. For example, on averag
e, ISS restores the CIFAR-10 model accuracy to 81.73%, surpassing the previous b
est preprocessing-based countermeasures by 37.97% absolute. ISS also (slightly)
outperforms adversarial training and has higher generalizability to unseen pertu
rbation norms and also higher efficiency. Our investigation reveals that the pro
perty of PAP perturbations depends on the type of surrogate model used for poiso
n generation, and it explains why a specific ISS compression yields the best per
formance for a specific type of PAP perturbation. We further test stronger, adap
tive poisoning, and show it falls short of being an ideal defense against ISS. O
verall, our results demonstrate the importance of considering various (simple) c
ountermeasures to ensure the meaningfulness of analysis carried out during the d
evelopment of availability poisons.
****

Which Invariance Should We Transfer? A Causal Minimax Learning Approach

Mingzhou Liu, Xiangyu Zheng, Xinwei Sun, Fang Fang, Yizhou Wang

A major barrier to deploying current machine learning models lies in their non-r
eliability to dataset shifts. To resolve this problem, most existing studies att
empted to transfer stable information to unseen environments. Particularly, inde
pendent causal mechanisms-based methods proposed to remove mutable causal mechan
isms via the do-operator. Compared to previous methods, the obtained stable pred
ictors are more effective in identifying stable information. However, a key ques
tion remains: which subset of this whole stable information should the model tra
nsfer, in order to achieve optimal generalization ability? To answer this questi
on, we present a comprehensive minimax analysis from a causal perspective. Speci
fically, we first provide a graphical condition for the whole stable set to be o
ptimal. When this condition fails, we surprisingly find with an example that thi
s whole stable set, although can fully exploit stable information, is not the op
timal one to transfer. To identify the optimal subset under this case, we propos
e to estimate the worst-case risk with a novel optimization scheme over the inte
rvention functions on mutable causal mechanisms. We then propose an efficient al
gorithm to search for the subset with minimal worst-case risk, based on a newly
defined equivalence relation between stable subsets. Compared to the exponential
 cost of exhaustively searching over all subsets, our searching strategy enjoys
a polynomial complexity. The effectiveness and efficiency of our methods are dem
onstrated on synthetic data and the diagnosis of Alzheimer’s disease.
****

Unsupervised Out-of-Distribution Detection with Diffusion Inpainting

Zhenzhen Liu, Jin Peng Zhou, Yufan Wang, Kilian Q Weinberger
Unsupervised out-of-distribution detection (OOD) seeks to identify out-of-domain data by learning only from unlabeled in-domain data. We present a novel approach for this task – Lift, Map, Detect (LMD) – that leverages recent advancement in diffusion models. Diffusion models are one type of generative models. At their core, they learn an iterative denoising process that gradually maps a noisy image closer to their training manifolds. LMD leverages this intuition for OOD detection. Specifically, LMD lifts an image off its original manifold by corrupting it, and maps it towards the in-domain manifold with a diffusion model. For an OOD image, the mapped image would have a large distance away from its original manifold, and LMD would identify it as OOD accordingly. We show through extensive experiments that LMD achieves competitive performance across a broad variety of datasets. Code can be found at https://github.com/zhenzhel/lift_map_detect.
****

NA^{2}Q: Neural Attention Additive Model for Interpretable Multi-Agent Q-Learning
Zichuan Liu, Yuanyang Zhu, Chunlin Chen
Value decomposition is widely used in cooperative multi-agent reinforcement learning, however, its implicit credit assignment mechanism is not yet fully understood due to black-box networks. In this work, we study an interpretable value decomposition framework via the family of generalized additive models. We present a novel method, named Neural Attention Additive Q-learning (NA^2Q), providing inherent intelligibility of collaboration behavior. NA^2Q can explicitly factorize the optimal joint policy induced by enriching shape functions to model all possible coalition of agents into individual policies. Moreover, we construct the identity semantics to promote estimating credits together with the global state and individual value functions, where local semantic masks help us diagnose whether each agent captures the relevant-task information. Extensive experiments show that NA^2Q consistently achieves superior performance compared to different state-of-the-art methods on all challenging tasks, while yielding human-like interpretability.
****

Contextual Combinatorial Bandits with Probabilistically Triggered Arms
Xutong Liu, Jinhang Zuo, Siwei Wang, John C.S. Lui, Mohammad Hajiesmaili, Adam Wierman, Wei Chen
We study contextual combinatorial bandits with probabilistically triggered arms (C^{2}MAB-T) under a variety of smoothness conditions that capture a wide range of applications, such as contextual cascading bandits and contextual influence maximization bandits. Under the triggering probability modulated (TPM) condition, we devise the C^2-UCB-T algorithm and propose a novel analysis that achieves an $\tilde{O}(d\sqrt{KT})$ regret bound, removing a potentially exponentially large factor $O(1/p_{\min})$, where d is the dimension of contexts, $p_{\min}$ is the minimum positive probability that any arm can be triggered, and batch-size K is the maximum number of arms that can be triggered per round. Under the variance modulated (VM) or triggering probability and variance modulated (TPVM) conditions, we propose a new variance-adaptive algorithm VAC^2-UCB and derive a regret bound $\tilde{O}(d\sqrt{T})$, which is independent of the batch-size K. As a valuable by-product, our analysis technique and variance-adaptive algorithm can be applied to the CMAB-T and C^{2}MAB setting, improving existing results there as well. We also include experiments that demonstrate the improved performance of our algorithms compared with benchmark algorithms on synthetic and real-world datasets.
****

Flipping Coins to Estimate Pseudocounts for Exploration in Reinforcement Learning
Sam Lobel, Akhil Bagaria, George Konidaris
We propose a new method for count-based exploration in high-dimensional state spaces. Unlike previous work which relies on density models, we show that counts can be derived by averaging samples from the Rademacher distribution (or coin flips). This insight is used to set up a simple supervised learning objective which

, when optimized, yields a state’s visitation count. We show that our method is significantly more effective at deducing ground-truth visitation counts than previous work; when used as an exploration bonus for a model-free reinforcement learning algorithm, it outperforms existing approaches on most of 9 challenging exploration tasks, including the Atari game Montezuma’s Revenge.
****
Multi-Symmetry Ensembles: Improving Diversity and Generalization via Opposing Symmetries
Charlotte Loh, Seungwook Han, Shivchander Sudalairaj, Rumen Dangovski, Kai Xu, Florian Wenzel, Marin Soljacic, Akash Srivastava
Deep ensembles (DE) have been successful in improving model performance by learning diverse members via the stochasticity of random initialization. While recent works have attempted to promote further diversity in DE via hyperparameters or regularizing loss functions, these methods primarily still rely on a stochastic approach to explore the hypothesis space. In this work, we present Multi-Symmetry Ensembles (MSE), a framework for constructing diverse ensembles by capturing the multiplicity of hypotheses along symmetry axes, which explore the hypothesis space beyond stochastic perturbations of model weights and hyperparameters. We leverage recent advances in contrastive representation learning to create models that separately capture opposing hypotheses of invariant and equivariant functional classes and present a simple ensembling approach to efficiently combine appropriate hypotheses for a given task. We show that MSE effectively captures the multiplicity of conflicting hypotheses that is often required in large, diverse datasets like ImageNet. As a result of their inherent diversity, MSE improves classification performance, uncertainty quantification, and generalization across a series of transfer tasks. Our code is available at https://github.com/clott3/multi-sym-ensem
****
The Flan Collection: Designing Data and Methods for Effective Instruction Tuning
Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, Adam Roberts
We study the design decision of publicly available instruction tuning methods, by reproducing and breaking down the development of Flan 2022 (Chung et al., 2022). Through careful ablation studies on the Flan Collection of tasks and methods, we tease apart the effect of design decisions which enable Flan-T5 to outperform prior work by 3-17% across evaluation settings. We find task balancing and enrichment techniques are overlooked but critical to effective instruction tuning, and in particular, training with mixed prompt settings (zero-shot, few-shot, chain-of-thought) actually yields equivalent or stronger (2%) performance in all settings. In further experiments we show Flan-T5 requires less finetuning to converge higher and faster than T5 on single downstream tasks – motivating instruction-tuned models as more computationally-efficient starting checkpoints for new tasks. Finally, to accelerate research on instruction tuning, we make the Flan 2022 collection of datasets, templates, and methods publicly available.
****
Dataset Distillation with Convexified Implicit Gradients
Noel Loo, Ramin Hasani, Mathias Lechner, Daniela Rus
We propose a new dataset distillation algorithm using reparameterization and convexification of implicit gradients (RCIG), that substantially improves the state-of-the-art. To this end, we first formulate dataset distillation as a bi-level optimization problem. Then, we show how implicit gradients can be effectively used to compute meta-gradient updates. We further equip the algorithm with a convexified approximation that corresponds to learning on top of a frozen finite-width neural tangent kernel. Finally, we improve bias in implicit gradients by parameterizing the neural network to enable analytical computation of final-layer parameters given the body parameters. RCIG establishes the new state-of-the-art on a diverse series of dataset distillation tasks. Notably, with one image per class, on resized ImageNet, RCIG sees on average a 108% improvement over the previous state-of-the-art distillation algorithm. Similarly, we observed a 66% gain over SOTA on Tiny-ImageNet and 37% on CIFAR-100.

****
Reflected Diffusion Models
Aaron Lou, Stefano Ermon
Score-based diffusion models learn to reverse a stochastic differential equation that maps data to noise. However, for complex tasks, numerical error can compound and result in highly unnatural samples. Previous work mitigates this drift with thresholding, which projects to the natural data domain (such as pixel space for images) after each diffusion step, but this leads to a mismatch between the training and generative processes. To incorporate data constraints in a principled manner, we present Reflected Diffusion Models, which instead reverse a reflected stochastic differential equation evolving on the support of the data. Our approach learns the perturbed score function through a generalized score matching loss and extends key components of standard diffusion models including diffusion guidance, likelihood-based training, and ODE sampling. We also bridge the theoretical gap with thresholding: such schemes are just discretizations of reflected SDEs. On standard image benchmarks, our method is competitive with or surpasses the state of the art without architectural modifications and, for classifier-free guidance, our approach enables fast exact sampling with ODEs and produces more faithful samples under high guidance weight.
****
Never mind the metrics---what about the uncertainty? Visualising binary confusion matrix metric distributions to put performance in perspective
David Lovell, Dimity Miller, Jaiden Capra, Andrew P. Bradley
There are strong incentives to build classification systems that show outstanding performance on various datasets and benchmarks. This can encourage a narrow focus on models and the performance metrics used to evaluate and compare them—resulting in a growing body of literature to evaluate and compare metrics. This paper strives for a more balanced perspective on binary classifier performance metrics by showing how uncertainty in these metrics can easily eclipse differences in empirical performance. We emphasise the discrete nature of confusion matrices and show how they can be well represented in a 3D lattice whose cross-sections form the space of receiver operating characteristic (ROC) curves. We develop novel interactive visualisations of performance metric contours within (and beyond) ROC space, showing the discrete probability mass functions of true and false positive rates and how these relate to performance metric distributions. We aim to raise awareness of the substantial uncertainty in performance metric estimates that can arise when classifiers are evaluated on empirical datasets and benchmarks, and that performance claims should be tempered by this understanding.
****
Bilevel Optimization with Coupled Decision-Dependent Distributions
Songtao Lu
Bilevel optimization has gained significant popularity in recent years due to its ability to formulate various machine learning problems. For instance, in meta-learning, the upper-level (UL) problem offers a good initialization for the lower-level (LL) model to facilitate adaptation. However, the decision variables can impact data features and outcomes, leading to the phenomenon known as performativity. In this work, we investigate the inclusion of decision-dependent distributions in bilevel optimization. Specifically, we consider the scenarios where the UL data distribution depends on the LL optimization variable, and the LL data distribution also depends on the UL decision variable. We first establish sufficient conditions for the existence of performatively stable (PS) solutions in this class of bilevel problems. Also, we propose efficient stochastic algorithms to find the PS point with theoretical convergence rate analysis and discuss the theoretical optimality of the obtained solution. Our theoretical analysis is corroborated through a series of numerical experiments, wherein we evaluate the performance of the bilevel performative prediction algorithms alongside non-performative counterparts in the context of meta strategic learning problems.
****
Two-Scale Gradient Descent Ascent Dynamics Finds Mixed Nash Equilibria of Continuous Games: A Mean-Field Perspective

Yulong Lu
Finding the mixed Nash equilibria (MNE) of a two-player zero sum continuous game is an important and challenging problem in machine learning. A canonical algorithm to finding the MNE is the noisy gradient descent ascent method which in the infinite particle limit gives rise to the Mean-Field Gradient Descent Ascent (GDA) dynamics on the space of probability measures. In this paper, we first study the convergence of a two-scale Mean-Field GDA dynamics for finding the MNE of the entropy-regularized objective. More precisely we show that for each finite temperature (or regularization parameter), the two-scale Mean-Field GDA with a suitable finite scale ratio converges exponentially to the unique MNE without assuming the convexity or concavity of the interaction potential. The key ingredient of our proof lies in the construction of new Lyapunov functions that dissipate exponentially along the Mean-Field GDA. We further study the simulated annealing of the Mean-Field GDA dynamics. We show that with a temperature schedule that decays logarithmically in time the annealed Mean-Field GDA converges to the MNE of the original unregularized objective.
****

STEP: Learning N:M Structured Sparsity Masks from Scratch with Precondition
Yucheng Lu, Shivani Agrawal, Suvinay Subramanian, Oleg Rybakov, Christopher De Sa, Amir Yazdanbakhsh
Recent innovations on hardware (e.g. Nvidia A100) have motivated learning N:M structured sparsity masks from scratch for fast model inference. However, state-of-the-art learning recipes in this regime (e.g. SR-STE) are proposed for non-adaptive optimizers like momentum SGD, while incurring non-trivial accuracy drop for Adam-trained models like attention-based LLMs. In this paper, we first demonstrate such gap origins from poorly estimated second moment (i.e. variance) in Adam states given by the masked weights. We conjecture that learning N:M masks with Adam should take the critical regime of variance estimation into account. In light of this, we propose STEP, an Adam-aware recipe that learns N:M masks with two phases: first, STEP calculates a reliable variance estimate (precondition phase) and subsequently, the variance remains fixed and is used as a precondition to learn N:M masks (mask-learning phase). STEP automatically identifies the switching point of two phases by dynamically sampling variance changes over the training trajectory and testing the sample concentration. Empirically, we evaluate STEP and other baselines such as ASP and SR-STE on multiple tasks including CIFAR classification, machine translation and LLM fine-tuning (BERT-Base, GPT-2). We show STEP mitigates the accuracy drop of baseline recipes and is robust to aggressive structured sparsity ratios.
****

Contrastive Energy Prediction for Exact Energy-Guided Diffusion Sampling in Offline Reinforcement Learning
Cheng Lu, Huayu Chen, Jianfei Chen, Hang Su, Chongxuan Li, Jun Zhu
Guided sampling is a vital approach for applying diffusion models in real-world tasks that embeds human-defined guidance during the sampling procedure. This paper considers a general setting where the guidance is defined by an (unnormalized) energy function. The main challenge for this setting is that the intermediate guidance during the diffusion sampling procedure, which is jointly defined by the sampling distribution and the energy function, is unknown and is hard to estimate. To address this challenge, we propose an exact formulation of the intermediate guidance as well as a novel training objective named contrastive energy prediction (CEP) to learn the exact guidance. Our method is guaranteed to converge to the exact guidance under unlimited model capacity and data samples, while previous methods can not. We demonstrate the effectiveness of our method by applying it to offline reinforcement learning (RL). Extensive experiments on D4RL benchmarks demonstrate that our method outperforms existing state-of-the-art algorithms. We also provide some examples of applying CEP for image synthesis to demonstrate the scalability of CEP on high-dimensional data.
****

Exploring the Limits of Model-Targeted Indiscriminate Data Poisoning Attacks
Yiwei Lu, Gautam Kamath, Yaoliang Yu

Indiscriminate data poisoning attacks aim to decrease a model’s test accuracy by injecting a small amount of corrupted training data. Despite significant interest, existing attacks remain relatively ineffective against modern machine learning (ML) architectures. In this work, we introduce the notion of model poisoning reachability as a technical tool to explore the intrinsic limits of data poisoning attacks towards target parameters (i.e., model-targeted attacks). We derive an easily computable threshold to establish and quantify a surprising phase transition phenomenon among popular ML models: data poisoning attacks can achieve certain target parameters only when the poisoning ratio exceeds our threshold. Building on existing parameter corruption attacks and refining the Gradient Canceling attack, we perform extensive experiments to confirm our theoretical findings, test the predictability of our transition threshold, and significantly improve existing indiscriminate data poisoning baselines over a range of datasets and models. Our work highlights the critical role played by the poisoning ratio, and sheds new insights on existing empirical results, attacks and mitigation strategies in data poisoning.
****

QAS-Bench: Rethinking Quantum Architecture Search and A Benchmark
Xudong Lu, Kaisen Pan, Ge Yan, Jiaming Shan, Wenjie Wu, Junchi Yan
Automatic quantum architecture search (QAS) has been widely studied across disciplines with different implications. In this paper, beyond a particular domain, we formulate the QAS problem into two basic (and relatively even ideal) tasks: i) arbitrary quantum circuit (QC) regeneration given a target QC; ii) approximating an arbitrary unitary (oracle). The latter can be connected to the setting of various quantum machine learning tasks and other QAS applications. Based on these two tasks, we generate a public QAS benchmark including 900 random QCs and 400 random unitary matrices which is still missing in the literature. We evaluate six baseline algorithms including brute force search, simulated annealing, genetic algorithm, reinforcement learning, hybrid algorithm, and differentiable algorithm as part of our benchmark. One characteristic of our proposed evaluation protocol on the basic tasks is that it deprives the domain-specific designs and techniques as used in existing QAS literature, making a unified evaluation possible and focusing on the vanilla search methods themselves without coupling with domain prior. In fact, the unitary approximation task could be algorithmically more difficult than the specific problems as it needs to explore the whole matrix space to fit the unitary. While specific tasks often only need to fit a partial observation of the unitary as the objective for search. Data and code are available at https://github.com/Lucky-Lance/QAS-Bench.
****

Learning Dense Correspondences between Photos and Sketches
Xuanchen Lu, Xiaolong Wang, Judith E Fan
Humans effortlessly grasp the connection between sketches and real-world objects, even when these sketches are far from realistic. Moreover, human sketch understanding goes beyond categorization – critically, it also entails understanding how individual elements within a sketch correspond to parts of the physical world it represents. What are the computational ingredients needed to support this ability? Towards answering this question, we make two contributions: first, we introduce a new sketch-photo correspondence benchmark, PSC6k, containing 150K annotations of 6250 sketch-photo pairs across 125 object categories, augmenting the existing Sketchy dataset with fine-grained correspondence metadata. Second, we propose a self-supervised method for learning dense correspondences between sketch-photo pairs, building upon recent advances in correspondence learning for pairs of photos. Our model uses a spatial transformer network to estimate the warp flow between latent representations of a sketch and photo extracted by a contrastive learning-based ConvNet backbone. We found that this approach outperformed several strong baselines and produced predictions that were quantitatively consistent with other warp-based methods. However, our benchmark also revealed systematic differences between predictions of the suite of models we tested and those of humans. Taken together, our work suggests a promising path towards developing artificial systems that achieve more human-like understanding of visual images at

different levels of abstraction. Project page: https://photo-sketch-correspondence.github.io

****

Adversarial Cheap Talk

Chris Lu, Timon Willi, Alistair Letcher, Jakob Nicolaus Foerster

Adversarial attacks in reinforcement learning (RL) often assume highly-privileged access to the victim's parameters, environment, or data. Instead, this paper proposes a novel adversarial setting called a Cheap Talk MDP in which an Adversary can merely append deterministic messages to the Victim's observation, resulting in a minimal range of influence. The Adversary cannot occlude ground truth, influence underlying environment dynamics or reward signals, introduce non-stationarity, add stochasticity, see the Victim's actions, or access their parameters. Additionally, we present a simple meta-learning algorithm called Adversarial Cheap Talk (ACT) to train Adversaries in this setting. We demonstrate that an Adversary trained with ACT can still significantly influence the Victim's training and testing performance, despite the highly constrained setting. Affecting train-time performance reveals a new attack vector and provides insight into the success and failure modes of existing RL algorithms. More specifically, we show that an ACT Adversary is capable of harming performance by interfering with the learner's function approximation, or instead helping the Victim's performance by outputting useful features. Finally, we show that an ACT Adversary can manipulate messages during train-time to directly and arbitrarily control the Victim at test-time.

****

Federated Conformal Predictors for Distributed Uncertainty Quantification

Charles Lu, Yaodong Yu, Sai Praneeth Karimireddy, Michael Jordan, Ramesh Raskar

Conformal prediction is emerging as a popular paradigm for providing rigorous uncertainty quantification in machine learning since it can be easily applied as a post-processing step to already trained models. In this paper, we extend conformal prediction to the federated learning setting. The main challenge we face is data heterogeneity across the clients – this violates the fundamental tenet of exchangeability required for conformal prediction. We propose a weaker notion of partial exchangeability, better suited to the FL setting, and use it to develop the Federated Conformal Prediction (FCP) framework. We show FCP enjoys rigorous theoretical guarantees and excellent empirical performance on several computer vision and medical imaging datasets. Our results demonstrate a practical approach to incorporating meaningful uncertainty quantification in distributed and heterogeneous environments. We provide code used in our experiments https://github.com/clu5/federated-conformal.

****

Mechanistic Mode Connectivity

Ekdeep Singh Lubana, Eric J Bigelow, Robert P. Dick, David Krueger, Hidenori Tanaka

We study neural network loss landscapes through the lens of mode connectivity, the observation that minimizers of neural networks retrieved via training on a dataset are connected via simple paths of low loss. Specifically, we ask the following question: are minimizers that rely on different mechanisms for making their predictions connected via simple paths of low loss? We provide a definition of mechanistic similarity as shared invariances to input transformations and demonstrate that lack of linear connectivity between two models implies they use dissimilar mechanisms for making their predictions. Relevant to practice, this result helps us demonstrate that naive fine-tuning on a downstream dataset can fail to alter a model's mechanisms, e.g., fine-tuning can fail to eliminate a model's reliance on spurious attributes. Our analysis also motivates a method for targeted alteration of a model's mechanisms, named connectivity-based fine-tuning (CBFT), which we analyze using several synthetic datasets for the task of reducing a model's reliance on spurious attributes.

****

A Unifying Framework to the Analysis of Interaction Methods using Synergy Functions

Daniel Lundstrom, Meisam Razaviyayn
Deep learning has revolutionized many areas of machine learning, from computer v ision to natural language processing, but these high-performance models are gene rally “black box." Explaining such models would improve transparency and trust i n AI-powered decision making and is necessary for understanding other practical needs such as robustness and fairness. A popular means of enhancing model transp arency is to quantify how individual inputs contribute to model outputs (called attributions) and the magnitude of interactions between groups of inputs. A grow ing number of these methods import concepts and results from game theory to prod uce attributions and interactions. This work presents a unifying framework for g ame-theory-inspired attribution and k^th-order interaction methods. We show that, given modest assumptions, a unique full account of interactions betwe en features, called synergies, is possible in the continuous input setting. We i dentify how various methods are characterized by their policy of distributing sy nergies. We establish that gradient-based methods are characterized by their act ions on monomials, a type of synergy function, and introduce unique gradient-bas ed methods. We show that the combination of various criteria uniquely defines th e attribution/interaction methods. Thus, the community needs to identify goals a nd contexts when developing and employing attribution and interaction methods.
****

SegCLIP: Patch Aggregation with Learnable Centers for Open-Vocabulary Semantic S egmentation
Huaishao Luo, Junwei Bao, Youzheng Wu, Xiaodong He, Tianrui Li
Recently, the contrastive language-image pre-training, e.g., CLIP, has demonstra ted promising results on various downstream tasks. The pre-trained model can cap ture enriched visual concepts for images by learning from a large scale of text- image data. However, transferring the learned visual knowledge to open-vocabular y semantic segmentation is still under-explored. In this paper, we propose a CLI P-based model named SegCLIP for the topic of open-vocabulary segmentation in an annotation-free manner. The SegCLIP achieves segmentation based on ViT and the m ain idea is to gather patches with learnable centers to semantic regions through training on text-image pairs. The gathering operation can dynamically capture t he semantic groups, which can be used to generate the final segmentation results . We further propose a reconstruction loss on masked patches and a superpixel-ba sed KL loss with pseudo-labels to enhance the visual representation. Experimenta l results show that our model achieves comparable or superior segmentation accur acy on the PASCAL VOC 2012 (+0.3% mIoU), PASCAL Context (+2.3% mIoU), and COCO (+2.2% mIoU) compared with baselines. We release the code at https://github.com/A rrowLuo/SegCLIP.
****

Image Restoration with Mean-Reverting Stochastic Differential Equations
Ziwei Luo, Fredrik K. Gustafsson, Zheng Zhao, Jens Sjölund, Thomas B. Schön
This paper presents a stochastic differential equation (SDE) approach for genera l-purpose image restoration. The key construction consists in a mean-reverting S DE that transforms a high-quality image into a degraded counterpart as a mean st ate with fixed Gaussian noise. Then, by simulating the corresponding reverse-tim e SDE, we are able to restore the origin of the low-quality image without relyin g on any task-specific prior knowledge. Crucially, the proposed mean-reverting S DE has a closed-form solution, allowing us to compute the ground truth time-depe ndent score and learn it with a neural network. Moreover, we propose a maximum l ikelihood objective to learn an optimal reverse trajectory that stabilizes the t raining and improves the restoration results. The experiments show that our prop osed method achieves highly competitive performance in quantitative comparisons on image deraining, deblurring, and denoising, setting a new state-of-the-art on two deraining datasets. Finally, the general applicability of our approach is f urther demonstrated via qualitative results on image super-resolution, inpaintin g, and dehazing. Code is available at https://github.com/Algolzw/image-restorati on-sde.
****

Dimensionality Reduction for General KDE Mode Finding

Xinyu Luo, Christopher Musco, Cas Widdershoven
Finding the mode of a high dimensional probability distribution $\mathcal{D}$ is a fundamental algorithmic problem in statistics and data analysis. There has been particular interest in efficient methods for solving the problem when $\mathcal{D}$ is represented as a mixture model or kernel density estimate, although few algorithmic results with worst-case approximation and runtime guarantees are known. In this work, we significantly generalize a result of (LeeLiMusco:2021) on mode approximation for Gaussian mixture models. We develop randomized dimensionality reduction methods for mixtures involving a broader class of kernels, including the popular logistic, sigmoid, and generalized Gaussian kernels. As in Lee et al.'s work, our dimensionality reduction results yield quasi-polynomial algorithms for mode finding with multiplicative accuracy $(1-\epsilon)$ for any $\epsilon > 0$. Moreover, when combined with gradient descent, they yield efficient practical heuristics for the problem. In addition to our positive results, we prove a hardness result for box kernels, showing that there is no polynomial time algorithm for finding the mode of a kernel density estimate, unless $\mathit{P} = \mathit{NP}$. Obtaining similar hardness results for kernels used in practice (like Gaussian or logistic kernels) is an interesting future direction.
****

Iterative Approximate Cross-Validation
Yuetian Luo, Zhimei Ren, Rina Barber
Cross-validation (CV) is one of the most popular tools for assessing and selecting predictive models. However, standard CV suffers from high computational cost when the number of folds is large. Recently, under the empirical risk minimization (ERM) framework, a line of works proposed efficient methods to approximate CV based on the solution of the ERM problem trained on the full dataset. However, in large-scale problems, it can be hard to obtain the exact solution of the ERM problem, either due to limited computational resources or due to early stopping as a way of preventing overfitting. In this paper, we propose a new paradigm to efficiently approximate CV when the ERM problem is solved via an iterative first-order algorithm, without running until convergence. Our new method extends existing guarantees for CV approximation to hold along the whole trajectory of the algorithm, including at convergence, thus generalizing existing CV approximation methods. Finally, we illustrate the accuracy and computational efficiency of our method through a range of empirical studies.
****

A Closer Look at Few-shot Classification Again
Xu Luo, Hao Wu, Ji Zhang, Lianli Gao, Jing Xu, Jingkuan Song
Few-shot classification consists of a training phase where a model is learned on a relatively large dataset and an adaptation phase where the learned model is adapted to previously-unseen tasks with limited labeled samples. In this paper, we empirically prove that the training algorithm and the adaptation algorithm can be completely disentangled, which allows algorithm analysis and design to be done individually for each phase. Our meta-analysis for each phase reveals several interesting insights that may help better understand key aspects of few-shot classification and connections with other fields such as visual representation learning and transfer learning. We hope the insights and research challenges revealed in this paper can inspire future work in related directions. Code and pre-trained models (in PyTorch) are available at https://github.com/Frankluox/CloserLookAgainFewShot.
****

HOPE: High-order Graph ODE For Modeling Interacting Dynamics
Xiao Luo, Jingyang Yuan, Zijie Huang, Huiyu Jiang, Yifang Qin, Wei Ju, Ming Zhang, Yizhou Sun
Leading graph ordinary differential equation (ODE) models have offered generalized strategies to model interacting multi-agent dynamical systems in a data-driven approach. They typically consist of a temporal graph encoder to get the initial states and a neural ODE-based generative model to model the evolution of dynamical systems. However, existing methods have severe deficiencies in capacity and efficiency due to the failure to model high-order correlations in long-term tem

poral trends. To tackle this, in this paper, we propose a novel model named High-order graph ODE (HOPE) for learning from dynamic interaction data, which can be naturally represented as a graph. It first adopts a twin graph encoder to initialize the latent state representations of nodes and edges, which consists of two branches to capture spatio-temporal correlations in complementary manners. More importantly, our HOPE utilizes a second-order graph ODE function which models the dynamics for both nodes and edges in the latent space respectively, which enables efficient learning of long-term dependencies from complex dynamical systems. Experiment results on a variety of datasets demonstrate both the effectiveness and efficiency of our proposed method.

****

Stabilizing GANs' Training with Brownian Motion Controller

Tianjiao Luo, Ziyu Zhu, Jianfei Chen, Jun Zhu

The training process of generative adversarial networks (GANs) is unstable and does not converge globally. In this paper, we examine the stability of GANs from the perspective of control theory and propose a universal higher-order noise-based controller called Brownian Motion Controller (BMC). Starting with the prototypical case of Dirac-GANs, we design a BMC to retrieve precisely the same but reachable optimal equilibrium. We theoretically prove that the training process of DiracGANs-BMC is globally exponential stable and derive bounds on the rate of convergence. Then we extend our BMC to normal GANs and provide implementation instructions on GANs-BMC. Our experiments show that our GANs-BMC effectively stabilizes GANs' training under StyleGANv2-ada frameworks with a faster rate of convergence, a smaller range of oscillation, and better performance in terms of FID score.

****

OCD: Learning to Overfit with Conditional Diffusion Models

Shahar Lutati, Lior Wolf

We present a dynamic model in which the weights are conditioned on an input sample x and are learned to match those that would be obtained by finetuning a base model on x and its label y. This mapping between an input sample and network weights is approximated by a denoising diffusion model. The diffusion model we employ focuses on modifying a single layer of the base model and is conditioned on the input, activations, and output of this layer. Since the diffusion model is stochastic in nature, multiple initializations generate different networks, forming an ensemble, which leads to further improvements. Our experiments demonstrate the wide applicability of the method for image classification, 3D reconstruction, tabular data, speech separation, and natural language processing.

****

DiscoBAX: Discovery of optimal intervention sets in genomic experiment design

Clare Lyle, Arash Mehrjou, Pascal Notin, Andrew Jesson, Stefan Bauer, Yarin Gal, Patrick Schwab

The discovery of therapeutics to treat genetically-driven pathologies relies on identifying genes involved in the underlying disease mechanism. Existing approaches search over the billions of potential interventions to maximize the expected influence on the target phenotype. However, to reduce the risk of failure in future stages of trials, practical experiment design aims to find a set of interventions that maximally change a target phenotype via diverse mechanisms. We propose DiscoBAX - a sample-efficient method for maximizing the rate of significant discoveries per experiment while simultaneously probing for a wide range of diverse mechanisms during a genomic experiment campaign. We provide theoretical guarantees of optimality under standard assumptions, and conduct a comprehensive experimental evaluation covering both synthetic as well as real-world experimental design tasks. DiscoBAX outperforms existing state-of-the-art methods for experimental design, selecting effective and diverse perturbations in biological systems.

****

Understanding Plasticity in Neural Networks

Clare Lyle, Zeyu Zheng, Evgenii Nikishin, Bernardo Avila Pires, Razvan Pascanu, Will Dabney

Plasticity, the ability of a neural network to quickly change its predictions in response to new information, is essential for the adaptability and robustness of deep reinforcement learning systems. Deep neural networks are known to lose plasticity over the course of training even in relatively simple learning problems, but the mechanisms driving this phenomenon are still poorly understood. This paper conducts a systematic empirical analysis into plasticity loss, with the goal of understanding the phenomenon mechanistically in order to guide the future development of targeted solutions. We find that loss of plasticity is deeply connected to changes in the curvature of the loss landscape, but that it often occurs in the absence of saturated units. Based on this insight, we identify a number of parameterization and optimization design choices which enable networks to better preserve plasticity over the course of training. We validate the utility of these findings on larger-scale RL benchmarks in the Arcade Learning Environment.

****

Bandits with Knapsacks: Advice on Time-Varying Demands

Lixing Lyu, Wang Chi Cheung

We consider a non-stationary Bandits with Knapsack problem. The outcome distribution at each time is scaled by a non-stationary quantity that signifies changing demand volumes. Instead of studying settings with limited non-stationarity, we investigate how online predictions on the total demand volume Q allows us to improve our performance guarantees. We show that, without any prediction, any online algorithm incurs a linear-in-T regret. In contrast, with online predictions on Q, we propose an online algorithm that judiciously incorporates the predictions, and achieve regret bounds that depends on the accuracy of the predictions. These bounds are shown to be tight in settings when prediction accuracy improves across time. Our theoretical results are corroborated by our numerical findings.

****

Pairwise Ranking Losses of Click-Through Rates Prediction for Welfare Maximization in Ad Auctions

Boxiang Lyu, Zhe Feng, Zachary Robertson, Sanmi Koyejo

We study the design of loss functions for click-through rates (CTR) to optimize (social) welfare in advertising auctions. Existing works either only focus on CTR predictions without consideration of business objectives (e.g., welfare) in auctions or assume that the distribution over the participants' expected cost-per-impression (eCPM) is known a priori, then use various additional assumptions on the parametric form of the distribution to derive loss functions for predicting CTRs. In this work, we bring back the welfare objectives of ad auctions into CTR predictions and propose a novel weighted rankloss to train the CTR model. Compared to existing literature, our approach provides a provable guarantee on welfare but without assumptions on the eCPMs' distribution while also avoiding the intractability of naively applying existing learning-to-rank methods. Further, we propose a theoretically justifiable technique for calibrating the losses using labels generated from a teacher network, only assuming that the teacher network has bounded ℓ_2 generalization error. Finally, we demonstrate the advantages of the proposed loss on synthetic and real-world data.

****

Which Tricks are Important for Learning to Rank?

Ivan Lyzhin, Aleksei Ustimenko, Andrey Gulin, Liudmila Prokhorenkova

Nowadays, state-of-the-art learning-to-rank methods are based on gradient-boosted decision trees (GBDT). The most well-known algorithm is LambdaMART which was proposed more than a decade ago. Recently, several other GBDT-based ranking algorithms were proposed. In this paper, we thoroughly analyze these methods in a unified setup. In particular, we address the following questions. Is direct optimization of a smoothed ranking loss preferable over optimizing a convex surrogate? How to properly construct and smooth surrogate ranking losses? To address these questions, we compare LambdaMART with YetiRank and StochasticRank methods and their modifications. We also propose a simple improvement of the YetiRank approach that allows for optimizing specific ranking loss functions. As a result, we gai

n insights into learning-to-rank techniques and obtain a new state-of-the-art al
gorithm.
****
Learning Neural Constitutive Laws from Motion Observations for Generalizable PDE
 Dynamics
Pingchuan Ma, Peter Yichen Chen, Bolei Deng, Joshua B. Tenenbaum, Tao Du, Chuang
 Gan, Wojciech Matusik
We propose a hybrid neural network (NN) and PDE approach for learning generaliza
ble PDE dynamics from motion observations. Many NN approaches learn an end-to-en
d model that implicitly models both the governing PDE and constitutive models (o
r material models). Without explicit PDE knowledge, these approaches cannot guar
antee physical correctness and have limited generalizability. We argue that the
governing PDEs are often well-known and should be explicitly enforced rather tha
n learned. Instead, constitutive models are particularly suitable for learning d
ue to their data-fitting nature. To this end, we introduce a new framework terme
d "Neural Constitutive Laws" (NCLaw), which utilizes a network architecture that
 strictly guarantees standard constitutive priors, including rotation equivarian
ce and undeformed state equilibrium. We embed this network inside a differentiab
le simulation and train the model by minimizing a loss function based on the dif
ference between the simulation and the motion observation. We validate NCLaw on
various large-deformation dynamical systems, ranging from solids to fluids. Afte
r training on a single motion trajectory, our method generalizes to new geometri
es, initial/boundary conditions, temporal ranges, and even multi-physics systems
. On these extremely out-of-distribution generalization tasks, NCLaw is orders-o
f-magnitude more accurate than previous NN approaches. Real-world experiments de
monstrate our method’s ability to learn constitutive laws from videos.
****
LIV: Language-Image Representations and Rewards for Robotic Control
Yecheng Jason Ma, Vikash Kumar, Amy Zhang, Osbert Bastani, Dinesh Jayaraman
We present Language-Image Value learning (LIV), a unified objective for vision-l
anguage representation and reward learning from action-free videos with text ann
otations. Exploiting a novel connection between dual reinforcement learning and
mutual information contrastive learning, the LIV objective trains a multi-modal
representation that implicitly encodes a universal value function for tasks spec
ified as language or image goals. We use LIV to pre-train the first control-cent
ric vision-language representation from large human video datasets such as EpicK
itchen. Given only a language or image goal, the pre-trained LIV model can assig
n dense rewards to each frame in videos of unseen robots or humans attempting th
at task in unseen environments. Further, when some target domain-specific data i
s available, the same objective can be used to fine-tune and improve LIV and eve
n other pre-trained representations for robotic control and reward specification
 in that domain. In our experiments on several simulated and real-world robot en
vironments, LIV models consistently outperform the best prior input state repres
entations for imitation learning, as well as reward specification methods for po
licy synthesis. Our results validate the advantages of joint vision-language rep
resentation and reward learning within the unified, compact LIV framework.
****
Graph Inductive Biases in Transformers without Message Passing
Liheng Ma, Chen Lin, Derek Lim, Adriana Romero-Soriano, Puneet K. Dokania, Mark
Coates, Philip Torr, Ser-Nam Lim
Transformers for graph data are increasingly widely studied and successful in nu
merous learning tasks. Graph inductive biases are crucial for Graph Transformers
, and previous works incorporate them using message-passing modules and/or posit
ional encodings. However, Graph Transformers that use message-passing inherit kn
own issues of message-passing, and differ significantly from Transformers used i
n other domains, thus making transfer of research advances more difficult. On th
e other hand, Graph Transformers without message-passing often perform poorly on
 smaller datasets, where inductive biases are more crucial. To bridge this gap,
we propose the Graph Inductive bias Transformer (GRIT) — a new Graph Transformer
 that incorporates graph inductive biases without using message passing. GRIT is

based on several architectural changes that are each theoretically and empirically justified, including: learned relative positional encodings initialized with random walk probabilities, a flexible attention mechanism that updates node and node-pair representations, and injection of degree information in each layer. We prove that GRIT is expressive — it can express shortest path distances and various graph propagation matrices. GRIT achieves state-of-the-art empirical performance across a variety of graph datasets, thus showing the power that Graph Transformers without message-passing can deliver.
****
Learning Signed Distance Functions from Noisy 3D Point Clouds via Noise to Noise Mapping
Baorui Ma, Yu-Shen Liu, Zhizhong Han
Learning signed distance functions (SDFs) from 3D point clouds is an important task in 3D computer vision. However, without ground truth signed distances, point normals or clean point clouds, current methods still struggle from learning SDFs from noisy point clouds. To overcome this challenge, we propose to learn SDFs via a noise to noise mapping, which does not require any clean point cloud or ground truth supervision for training. Our novelty lies in the noise to noise mapping which can infer a highly accurate SDF of a single object or scene from its multiple or even single noisy point cloud observations. Our novel learning manner is supported by modern Lidar systems which capture multiple noisy observations per second. We achieve this by a novel loss which enables statistical reasoning on point clouds and maintains geometric consistency although point clouds are irregular, unordered and have no point correspondence among noisy observations. Our evaluation under the widely used benchmarks demonstrates our superiority over the state-of-the-art methods in surface reconstruction, point cloud denoising and upsampling. Our code, data, and pre-trained models are available at https://github.com/mabaorui/Noise2NoiseMapping/ .
****
Learning Intuitive Policies Using Action Features
Mingwei Ma, Jizhou Liu, Samuel Sokota, Max Kleiman-Weiner, Jakob Nicolaus Foerster
An unaddressed challenge in multi-agent coordination is to enable AI agents to exploit the semantic relationships between the features of actions and the features of observations. Humans take advantage of these relationships in highly intuitive ways. For instance, in the absence of a shared language, we might point to the object we desire or hold up our fingers to indicate how many objects we want. To address this challenge, we investigate the effect of network architecture on the propensity of learning algorithms to exploit these semantic relationships. Across a procedurally generated coordination task, we find that attention-based architectures that jointly process a featurized representation of observations and actions have a better inductive bias for learning intuitive policies. Through fine-grained evaluation and scenario analysis, we show that the resulting policies are human-interpretable. Moreover, such agents coordinate with people without training on any human data.
****
Over-parametrization via Lifting for Low-rank Matrix Sensing: Conversion of Spurious Solutions to Strict Saddle Points
Ziye Ma, Igor Molybog, Javad Lavaei, Somayeh Sojoudi
This paper studies the role of over-parametrization in solving non-convex optimization problems. The focus is on the important class of low-rank matrix sensing, where we propose an infinite hierarchy of non-convex problems via the lifting technique and the Burer-Monteiro factorization. This contrasts with the existing over-parametrization technique where the search rank is limited by the dimension of the matrix and it does not allow a rich over-parametrization of an arbitrary degree. We show that although the spurious solutions of the problem remain stationary points through the hierarchy, they will be transformed into strict saddle points (under some technical conditions) and can be escaped via local search methods. This is the first result in the literature showing that over-parametrization creates a negative curvature for escaping spurious solutions. We also derive

a bound on how much over-parametrization is requited to enable the elimination of spurious solutions.

****

Buying Information for Stochastic Optimization

Mingchen Ma, Christos Tzamos

Stochastic optimization is one of the central problems in Machine Learning and Theoretical Computer Science. In the standard model, the algorithm is given a fixed distribution known in advance. In practice though, one may acquire at a cost extra information to make better decisions. In this paper, we study how to buy information for stochastic optimization and formulate this question as an online learning problem. Assuming the learner has an oracle for the original optimization problem, we design a 2-competitive deterministic algorithm and a $e/(e-1)$-competitive randomized algorithm for buying information. We show that this ratio is tight as the problem is equivalent to a robust generalization of the ski-rental problem, which we call super-martingale stopping. We also consider an adaptive setting where the learner can choose to buy information after taking some actions for the underlying optimization problem. We focus on the classic optimization problem, Min-Sum Set Cover, where the goal is to quickly find an action that covers a given request drawn from a known distribution. We provide an 8-competitive algorithm running in polynomial time that chooses actions and decides when to buy information about the underlying request.

****

Generated Graph Detection

Yihan Ma, Zhikun Zhang, Ning Yu, Xinlei He, Michael Backes, Yun Shen, Yang Zhang

Graph generative models become increasingly effective for data distribution approximation and data augmentation. While they have aroused public concerns about their malicious misuses or misinformation broadcasts, just as what Deepfake visual and auditory media has been delivering to society. Hence it is essential to regulate the prevalence of generated graphs. To tackle this problem, we pioneer the formulation of the generated graph detection problem to distinguish generated graphs from real ones. We propose the first framework to systematically investigate a set of sophisticated models and their performance in four classification scenarios. Each scenario switches between seen and unseen datasets/generators during testing to get closer to real-world settings and progressively challenge the classifiers. Extensive experiments evidence that all the models are qualified for generated graph detection, with specific models having advantages in specific scenarios. Resulting from the validated generality and oblivion of the classifiers to unseen datasets/generators, we draw a safe conclusion that our solution can sustain for a decent while to curb generated graph misuses.

****

Calibrating Multimodal Learning

Huan Ma, Qingyang Zhang, Changqing Zhang, Bingzhe Wu, Huazhu Fu, Joey Tianyi Zhou, Qinghua Hu

Multimodal machine learning has achieved remarkable progress in a wide range of scenarios. However, the reliability of multimodal learning remains largely unexplored. In this paper, through extensive empirical studies, we identify current multimodal classification methods suffer from unreliable predictive confidence that tend to rely on partial modalities when estimating confidence. Specifically, we find that the confidence estimated by current models could even increase when some modalities are corrupted. To address the issue, we introduce an intuitive principle for multimodal learning, i.e., the confidence should not increase when one modality is removed. Accordingly, we propose a novel regularization technique, i.e., Calibrating Multimodal Learning (CML) regularization, to calibrate the predictive confidence of previous methods. This technique could be flexibly equipped by existing models and improve the performance in terms of confidence calibration, classification accuracy, and model robustness.

****

AutoCoreset: An Automatic Practical Coreset Construction Framework

Alaa Maalouf, Murad Tukan, Vladimir Braverman, Daniela Rus

A coreset is a small weighted subset of an input set that approximates its loss

function, for a given set of queries. Coresets became prevalent in machine learn ing as they have shown to be advantageous for many applications. Unfortunately, coresets are constructed in a problem-dependent manner, where for each problem, a new coreset construction algorithm is suggested, taking years to prove its cor rectness. Even the generic frameworks require additional (problem-dependent) com putations or proofs to be done by the user. Besides, many problems do not have (provable) small coresets, limiting their applicability. To this end, we suggest an automatic practical framework for constructing coresets, which requires (only) the input data and the desired cost function from the user, without the need f or any other task-related computation to be done by the user. To do so, we reduc e the problem of approximating a loss function to an instance of vector summatio n approximation, where the vectors we aim to sum are loss vectors of a specific subset of the queries, such that we aim to approximate the image of the function on this subset. We show that while this set is limited, the coreset is quite ge neral. An extensive experimental study on various machine learning applications is also conducted. Finally, we provide a “plug and play" style implementation, p roposing a user-friendly system that can be easily used to apply coresets for ma ny problems. We believe that these contributions enable future research and easi er use and applications of coresets.
****

Learning GFlowNets From Partial Episodes For Improved Convergence And Stability
Kanika Madan, Jarrid Rector-Brooks, Maksym Korablyov, Emmanuel Bengio, Moksh Jai n, Andrei Cristian Nica, Tom Bosc, Yoshua Bengio, Nikolay Malkin
Generative flow networks (GFlowNets) are a family of algorithms for training a s equential sampler of discrete objects under an unnormalized target density and h ave been successfully used for various probabilistic modeling tasks. Existing tr aining objectives for GFlowNets are either local to states or transitions, or pr opagate a reward signal over an entire sampling trajectory. We argue that these alternatives represent opposite ends of a gradient bias-variance tradeoff and pr opose a way to exploit this tradeoff to mitigate its harmful effects. Inspired b y the TD(λ) algorithm in reinforcement learning, we introduce subtraject ory balance or SubTB(λ), a GFlowNet training objective that can learn fr om partial action subsequences of varying lengths. We show that SubTB(λ) accelerates sampler convergence in previously studied and new environments and enables training GFlowNets in environments with longer action sequences and spar ser reward landscapes than what was possible before. We also perform a comparati ve analysis of stochastic gradient dynamics, shedding light on the bias-variance tradeoff in GFlowNet training and the advantages of subtrajectory balance.
****

Applied Online Algorithms with Heterogeneous Predictors
Jessica Maghakian, Russell Lee, Mohammad Hajiesmaili, Jian Li, Ramesh Sitaraman, Zhenhua Liu
For many application domains, the integration of machine learning (ML) models in to decision making is hindered by the poor explainability and theoretical guaran tees of black box models. Although the emerging area of algorithms with predicti ons offers a way to leverage ML while enjoying worst-case guarantees, existing w ork usually assumes access to only one predictor. We demonstrate how to more eff ectively utilize historical datasets and application domain knowledge by intenti onally using predictors of different quantities. By leveraging the heterogeneity in our predictors, we are able to achieve improved performance, explainability and computational efficiency over predictor-agnostic methods. Theoretical result s are supplemented by large-scale empirical evaluations with production data dem onstrating the success of our methods on optimization problems occurring in larg e distributed computing systems.
****

CSP: Self-Supervised Contrastive Spatial Pre-Training for Geospatial-Visual Repr esentations
Gengchen Mai, Ni Lao, Yutong He, Jiaming Song, Stefano Ermon
Geo-tagged images are publicly available in large quantities, whereas labels suc h as object classes are rather scarce and expensive to collect. Meanwhile, contr

astive learning has achieved tremendous success in various natural image and lan guage tasks with limited labeled data. However, existing methods fail to fully l everage geospatial information, which can be paramount to distinguishing objects that are visually similar. To directly leverage the abundant geospatial informa tion associated with images in pre-training, fine-tuning, and inference stages, we present Contrastive Spatial Pre-Training (CSP), a self-supervised learning fr amework for geo-tagged images. We use a dual-encoder to separately encode the im ages and their corresponding geo-locations, and use contrastive objectives to le arn effective location representations from images, which can be transferred to downstream supervised tasks such as image classification. Experiments show that CSP can improve model performance on both iNat2018 and fMoW datasets. Especially , on iNat2018, CSP significantly boosts the model performance with 10-34% relati ve improvement with various labeled training data sampling ratios.

****

Vertical Federated Graph Neural Network for Recommender System

Peihua Mai, Yan Pang

Conventional recommender systems are required to train the recommendation model using a centralized database. However, due to data privacy concerns, this is oft en impractical when multi-parties are involved in recommender system training. F ederated learning appears as an excellent solution to the data isolation and pri vacy problem. Recently, Graph neural network (GNN) is becoming a promising appro ach for federated recommender systems. However, a key challenge is to conduct em bedding propagation while preserving the privacy of the graph structure. Few stu dies have been conducted on the federated GNN-based recommender system. Our stud y proposes the first vertical federated GNN-based recommender system, called Ver FedGNN. We design a framework to transmit: (i) the summation of neighbor embeddi ngs using random projection, and (ii) gradients of public parameter perturbed by ternary quantization mechanism. Empirical studies show that VerFedGNN has compe titive prediction accuracy with existing privacy preserving GNN frameworks while enhanced privacy protection for users’ interaction information.

****

Can Neural Network Memorization Be Localized?

Pratyush Maini, Michael Curtis Mozer, Hanie Sedghi, Zachary Chase Lipton, J Zico Kolter, Chiyuan Zhang

Recent efforts at explaining the interplay of memorization and generalization in deep overparametrized networks have posited that neural networks memorize “hard ” examples in the final few layers of the model. Memorization refers to the abil ity to correctly predict on atypical examples of the training set. In this work, we show that rather than being confined to individual layers, memorization is a phenomenon confined to a small set of neurons in various layers of the model. F irst, via three experimental sources of converging evidence, we find that most l ayers are redundant for the memorization of examples and the layers that contrib ute to example memorization are, in general, not the final layers. The three sou rces are gradient accounting (measuring the contribution to the gradient norms f rom memorized and clean examples), layer rewinding (replacing specific model wei ghts of a converged model with previous training checkpoints), and retraining (t raining rewound layers only on clean examples). Second, we ask a more generic qu estion: can memorization be localized anywhere in a model? We discover that memo rization is often confined to a small number of neurons or channels (around 5) o f the model. Based on these insights we propose a new form of dropout—example-ti ed dropout that enables us to direct the memorization of examples to an aprior d etermined set of neurons. By dropping out these neurons, we are able to reduce t he accuracy on memorized examples from 100% to 3%, while also reducing the gener alization gap.

****

Fundamental Tradeoffs in Learning with Prior Information

Anirudha Majumdar

We seek to understand fundamental tradeoffs between the accuracy of prior inform ation that a learner has on a given problem and its learning performance. We int roduce the notion of prioritized risk, which differs from traditional notions of

minimax and Bayes risk by allowing us to study such fundamental tradeoffs in se
ttings where reality does not necessarily conform to the learner’s prior. We pre
sent a general reduction-based approach for extending classical minimax lower-bo
und techniques in order to lower bound the prioritized risk for statistical esti
mation problems. We also introduce a novel generalization of Fano’s inequality (
which may be of independent interest) for lower bounding the prioritized risk in
more general settings involving unbounded losses. We illustrate the ability of
our framework to provide insights into tradeoffs between prior information and l
earning performance for problems in estimation, regression, and reinforcement le
arning.
****

Additive Causal Bandits with Unknown Graph
Alan Malek, Virginia Aglietti, Silvia Chiappa
We explore algorithms to select actions in the causal bandit setting where the l
earner can choose to intervene on a set of random variables related by a causal
graph, and the learner sequentially chooses interventions and observes a sample
from the interventional distribution. The learner’s goal is to quickly find the
intervention, among all interventions on observable variables, that maximizes th
e expectation of an outcome variable. We depart from previous literature by assu
ming no knowledge of the causal graph except that latent confounders between the
outcome and its ancestors are not present. We first show that the unknown graph
problem can be exponentially hard in the parents of the outcome. To remedy this
, we adopt an additional additive assumption on the outcome which allows us to s
olve the problem by casting it as an additive combinatorial linear bandit proble
m with full-bandit feedback. We propose a novel action-elimination algorithm for
this setting, show how to apply this algorithm to the causal bandit problem, pr
ovide sample complexity bounds, and empirically validate our findings on a suite
of randomly generated causal models, effectively showing that one does not need
to explicitly learn the parents of the outcome to identify the best interventio
n.
****

Weighted Tallying Bandits: Overcoming Intractability via Repeated Exposure Optim
ality
Dhruv Malik, Conor Igoe, Yuanzhi Li, Aarti Singh
In human-interactive applications of online learning, a human’s preferences or a
bilities are often a function of the algorithm’s recent actions. Motivated by th
is, a significant line of work has formalized settings where an action’s loss is
a function of the number of times it was played in the prior m timesteps, whe
re m corresponds to a bound on human memory capacity. To more faithfully captu
re decay of human memory with time, we introduce the Weighted Tallying Bandit (W
TB), which generalizes this setting by requiring that an action’s loss is a func
tion of a weighted summation of the number of times it was played in the last $m
$ timesteps. WTB is intractable without further assumption. So we study it under
Repeated Exposure Optimality (REO), a condition requiring the existence of an a
ction that when repetitively played will eventually yield smaller loss than any
other action sequence. We study the minimization of complete policy regret (CPR)
, which is the strongest notion of regret, in WTB under REO. Since m is often
unknown, we only assume access to an upper bound M on m. We show that for pr
oblems with K actions and horizon T, a simple modification of the successive
elimination algorithm has $\mathcal{O} \left(\sqrt{KT} + (m+M)K \right)$ CPR.
Upto an additive (in lieu of mutliplicative) factor in $(m+M)K$, this recovers t
he classical guarantee for the far simpler stochastic multi-armed bandit with tr
aditional regret. We additionally show that in our setting, any algorithm will s
uffer additive CPR of $\Omega \left(mK + M \right)$, demonstrating our result i
s near optimal. Our method is computationally efficient, and we experimentally d
emonstrate its practicality and superiority over various baselines.
****

A Kernel-Based View of Language Model Fine-Tuning
Sadhika Malladi, Alexander Wettig, Dingli Yu, Danqi Chen, Sanjeev Arora
It has become standard to solve NLP tasks by fine-tuning pre-trained language mo

dels (LMs), especially in low-data settings. There is minimal theoretical unders tanding of empirical success, e.g., why fine-tuning a model with 10^8 or more parameters on a couple dozen training points does not result in overfitting. We investigate whether the Neural Tangent Kernel (NTK)—which originated as a model to study the gradient descent dynamics of infinitely wide networks with suitable random initialization—describes fine-tuning of pre-trained LMs. This study was inspired by the decent performance of NTK for computer vision tasks (Wei et al., 2022). We extend the NTK formalism to Adam and use Tensor Programs (Yang, 2020) to characterize conditions under which the NTK lens may describe fine-tuning up dates to pre-trained language models. Extensive experiments on 14 NLP tasks vali date our theory and show that formulating the downstream task as a masked word p rediction problem through prompting often induces kernel-based dynamics during f ine-tuning. Finally, we use this kernel view to propose an explanation for the s uccess of parameter-efficient subspace-based fine-tuning methods.

****

Performative Reinforcement Learning

Debmalya Mandal, Stelios Triantafyllou, Goran Radanovic

We introduce the framework of performative reinforcement learning where the poli cy chosen by the learner affects the underlying reward and transition dynamics o f the environment. Following the recent literature on performative prediction (P erdomo et al., 2020), we introduce the concept of performatively stable policy. We then consider a regularized version of the reinforcement learning problem and show that repeatedly optimizing this objective converges to a performatively st able policy under reasonable assumptions on the transition dynamics. Our proof u tilizes the dual perspective of the reinforcement learning problem and may be of independent interest in analyzing the convergence of other algorithms with deci sion-dependent environments. We then extend our results for the setting where th e learner just performs gradient ascent steps instead of fully optimizing the ob jective, and for the setting where the learner has access to a finite number of trajectories from the changed environment. For both the settings, we leverage th e dual formulation of performative reinforcement learning, and establish converg ence to a stable solution. Finally, through extensive experiments on a grid-worl d environment, we demonstrate the dependence of convergence on various parameter s e.g. regularization, smoothness, and the number of samples.

****

Differential Privacy has Bounded Impact on Fairness in Classification

Paul Mangold, Michaël Perrot, Aurélien Bellet, Marc Tommasi

We theoretically study the impact of differential privacy on fairness in classif ication. We prove that, given a class of models, popular group fairness measures are pointwise Lipschitz-continuous with respect to the parameters of the model. This result is a consequence of a more general statement on accuracy conditione d on an arbitrary event (such as membership to a sensitive group), which may be of independent interest. We use this Lipschitz property to prove a non-asymptoti c bound showing that, as the number of samples increases, the fairness level of private models gets closer to the one of their non-private counterparts. This bo und also highlights the importance of the confidence margin of a model on the di sparate impact of differential privacy.

****

Random Classification Noise does not defeat All Convex Potential Boosters Irresp ective of Model Choice

Yishay Mansour, Richard Nock, Robert Williamson

A landmark negative result of Long and Servedio has had a considerable impact on research and development in boosting algorithms, around the now famous tagline that "noise defeats all convex boosters". In this paper, we appeal to the half-c entury+ founding theory of losses for class probability estimation, an extension of Long and Servedio’s results and a new general convex booster to demonstrate that the source of their negative result is in fact the model class, linear sepa rators. Losses or algorithms are neither to blame. This leads us to a discussion on an otherwise praised aspect of ML, parameterisation.

****

H-Consistency Bounds for Pairwise Misranking Loss Surrogates
Anqi Mao, Mehryar Mohri, Yutao Zhong
We present a detailed study of H-consistency bounds for score-based ranking. T
hese are upper bounds on the target loss estimation error of a predictor in a hy
pothesis set H, expressed in terms of the surrogate loss estimation error of t
hat predictor. We will show that both in the general pairwise ranking scenario a
nd in the bipartite ranking scenario, there are no meaningful H-consistency bo
unds for most hypothesis sets used in practice including the family of linear mo
dels and that of the neural networks, which satisfy the equicontinuous property
with respect to the input. To come up with ranking surrogate losses with theoret
ical guarantees, we show that a natural solution consists of resorting to a pair
wise abstention loss in the general pairwise ranking scenario, and similarly, a
bipartite abstention loss in the bipartite ranking scenario, to abstain from mak
ing predictions at some limited cost c. For surrogate losses of these abstenti
on loss functions, we give a series of H-consistency bounds for both the famil
y of linear functions and that of neural networks with one hidden-layer. Our exp
erimental results illustrate the effectiveness of ranking with abstention.
****

Cross-Entropy Loss Functions: Theoretical Analysis and Applications
Anqi Mao, Mehryar Mohri, Yutao Zhong
Cross-entropy is a widely used loss function in applications. It coincides with
the logistic loss applied to the outputs of a neural network, when the softmax i
s used. But, what guarantees can we rely on when using cross-entropy as a surrog
ate loss? We present a theoretical analysis of a broad family of loss functions,
 comp-sum losses, that includes cross-entropy (or logistic loss), generalized cr
oss-entropy, the mean absolute error and other cross-entropy-like loss functions
. We give the first H-consistency bounds for these loss functions. These are n
on-asymptotic guarantees that upper bound the zero-one loss estimation error in
terms of the estimation error of a surrogate loss, for the specific hypothesis s
et H used. We further show that our bounds are tight. These bounds depend on q
uantities called minimizability gaps. To make them more explicit, we give a spec
ific analysis of these gaps for comp-sum losses. We also introduce a new family
of loss functions, smooth adversarial comp-sum losses, that are derived from the
ir comp-sum counterparts by adding in a related smooth term. We show that these
loss functions are beneficial in the adversarial setting by proving that they ad
mit H-consistency bounds. This leads to new adversarial robustness algorithms
that consist of minimizing a regularized smooth adversarial comp-sum loss. While
 our main purpose is a theoretical analysis, we also present an extensive empiri
cal analysis comparing comp-sum losses. We further report the results of a serie
s of experiments demonstrating that our adversarial robustness algorithms outper
form the current state-of-the-art, while also achieving a superior non-adversari
al accuracy.
****

Supported Trust Region Optimization for Offline Reinforcement Learning
Yixiu Mao, Hongchang Zhang, Chen Chen, Yi Xu, Xiangyang Ji
Offline reinforcement learning suffers from the out-of-distribution issue and ex
trapolation error. Most policy constraint methods regularize the density of the
trained policy towards the behavior policy, which is too restrictive in most cas
es. We propose Supported Trust Region optimization (STR) which performs trust re
gion policy optimization with the policy constrained within the support of the b
ehavior policy, enjoying the less restrictive support constraint. We show that,
when assuming no approximation and sampling error, STR guarantees strict policy
improvement until convergence to the optimal support-constrained policy in the d
ataset. Further with both errors incorporated, STR still guarantees safe policy
improvement for each step. Empirical results validate the theory of STR and demo
nstrate its state-of-the-art performance on MuJoCo locomotion domains and much m
ore challenging AntMaze domains.
****

Robust Perception through Equivariance
Chengzhi Mao, Lingyu Zhang, Abhishek Vaibhav Joshi, Junfeng Yang, Hao Wang, Carl

Vondrick

Deep networks for computer vision are not reliable when they encounter adversarial examples. In this paper, we introduce a framework that uses the dense intrinsic constraints in natural images to robustify inference. By introducing constraints at inference time, we can shift the burden of robustness from training to testing, thereby allowing the model to dynamically adjust to each individual image's unique and potentially novel characteristics at inference time. Our theoretical results show the importance of having dense constraints at inference time. In contrast to existing single-constraint methods, we propose to use equivariance, which naturally allows dense constraints at a fine-grained level in the feature space. Our empirical experiments show that restoring feature equivariance at inference time defends against worst-case adversarial perturbations. The method obtains improved adversarial robustness on four datasets (ImageNet, Cityscapes, PASCAL VOC, and MS-COCO) on image recognition, semantic segmentation, and instance segmentation tasks.

****

Reliable Measures of Spread in High Dimensional Latent Spaces

Anna Marbut, Katy Mckinney-Bock, Travis Wheeler

Understanding geometric properties of the latent spaces of natural language processing models allows the manipulation of these properties for improved performance on downstream tasks. One such property is the amount of data spread in a model's latent space, or how fully the available latent space is being used. We demonstrate that the commonly used measures of data spread, average cosine similarity and a partition function min/max ratio I(V), do not provide reliable metrics to compare the use of latent space across data distributions. We propose and examine six alternative measures of data spread, all of which improve over these current metrics when applied to seven synthetic data distributions. Of our proposed measures, we recommend one principal component-based measure and one entropy-based measure that provide reliable, relative measures of spread and can be used to compare models of different sizes and dimensionalities.

****

SRATTA: Sample Re-ATTribution Attack of Secure Aggregation in Federated Learning.

Tanguy Marchand, Regis Loeb, Ulysse Marteau-Ferey, Jean Ogier Du Terrail, Arthur Pignet

We consider a federated learning (FL) setting where a machine learning model with a fully connected first layer is trained between different clients and a central server using FedAvg, and where the aggregation step can be performed with secure aggregation (SA). We present SRATTA an attack relying only on aggregated models which, under realistic assumptions, (i) recovers data samples from the different clients, and (ii) groups data samples coming from the same client together. While sample recovery has already been explored in an FL setting, the ability to group samples per client, despite the use of SA, is novel. This poses a significant unforeseen security threat to FL and effectively breaks SA. We show that SRATTA is both theoretically grounded and can be used in practice on realistic models and datasets. We also propose counter-measures, and claim that clients should play an active role to guarantee their privacy during training.

****

Neuro-Symbolic Continual Learning: Knowledge, Reasoning Shortcuts and Concept Rehearsal

Emanuele Marconato, Gianpaolo Bontempo, Elisa Ficarra, Simone Calderara, Andrea Passerini, Stefano Teso

We introduce Neuro-Symbolic Continual Learning, where a model has to solve a sequence of neuro-symbolic tasks, that is, it has to map sub-symbolic inputs to high-level concepts and compute predictions by reasoning consistently with prior knowledge. Our key observation is that neuro-symbolic tasks, although different, often share concepts whose semantics remains stable over time. Traditional approaches fall short: existing continual strategies ignore knowledge altogether, while stock neuro-symbolic architectures suffer from catastrophic forgetting. We show that leveraging prior knowledge by combining neuro-symbolic architectures with

continual strategies does help avoid catastrophic forgetting, but also that doing so can yield models affected by reasoning shortcuts. These undermine the semantics of the acquired concepts, even when detailed prior knowledge is provided upfront and inference is exact, and in turn continual performance. To overcome these issues, we introduce COOL, a COncept-level cOntinual Learning strategy tailored for neuro-symbolic continual problems that acquires high-quality concepts and remembers them over time. Our experiments on three novel benchmarks highlights how COOL attains sustained high performance on neuro-symbolic continual learning tasks in which other strategies fail.

****

Evaluating Unsupervised Denoising Requires Unsupervised Metrics

Adria Marcos Morales, Matan Leibovich, Sreyas Mohan, Joshua Lawrence Vincent, Piyush Haluai, Mai Tan, Peter Crozier, Carlos Fernandez-Granda

Unsupervised denoising is a crucial challenge in real-world imaging applications. Unsupervised deep-learning methods have demonstrated impressive performance on benchmarks based on synthetic noise. However, no metrics exist to evaluate these methods in an unsupervised fashion. This is highly problematic for the many practical applications where ground-truth clean images are not available. In this work, we propose two novel metrics: the unsupervised mean squared error (MSE) and the unsupervised peak signal-to-noise ratio (PSNR), which are computed using only noisy data. We provide a theoretical analysis of these metrics, showing that they are asymptotically consistent estimators of the supervised MSE and PSNR. Controlled numerical experiments with synthetic noise confirm that they provide accurate approximations in practice. We validate our approach on real-world data from two imaging modalities: videos in raw format and transmission electron microscopy. Our results demonstrate that the proposed metrics enable unsupervised evaluation of denoising methods based exclusively on noisy data.

****

Regions of Reliability in the Evaluation of Multivariate Probabilistic Forecasts

Étienne Marcotte, Valentina Zantedeschi, Alexandre Drouin, Nicolas Chapados

Multivariate probabilistic time series forecasts are commonly evaluated via proper scoring rules, i.e., functions that are minimal in expectation for the ground-truth distribution. However, this property is not sufficient to guarantee good discrimination in the non-asymptotic regime. In this paper, we provide the first systematic finite-sample study of proper scoring rules for time series forecasting evaluation. Through a power analysis, we identify the “region of reliability” of a scoring rule, i.e., the set of practical conditions where it can be relied on to identify forecasting errors. We carry out our analysis on a comprehensive synthetic benchmark, specifically designed to test several key discrepancies between ground-truth and forecast distributions, and we gauge the generalizability of our findings to real-world tasks with an application to an electricity production problem. Our results reveal critical shortcomings in the evaluation of multivariate probabilistic forecasts as commonly performed in the literature.

****

Analyzing Diffusion as Serial Reproduction

Raja Marjieh, Ilia Sucholutsky, Thomas A Langlois, Nori Jacoby, Thomas L. Griffiths

Diffusion models are a class of generative models that learn to synthesize samples by inverting a diffusion process that gradually maps data into noise. While these models have enjoyed great success recently, a full theoretical understanding of their observed properties is still lacking, in particular, their weak sensitivity to the choice of noise family and the role of adequate scheduling of noise levels for good synthesis. By identifying a correspondence between diffusion models and a well-known paradigm in cognitive science known as serial reproduction, whereby human agents iteratively observe and reproduce stimuli from memory, we show how the aforementioned properties of diffusion models can be explained as a natural consequence of this correspondence. We then complement our theoretical analysis with simulations that exhibit these key features. Our work highlights how classic paradigms in cognitive science can shed light on state-of-the-art machine learning problems.

****
Quantized Distributed Training of Large Models with Convergence Guarantees
Ilia Markov, Adrian Vladu, Qi Guo, Dan Alistarh
Communication-reduction techniques are a popular way to improve scalability in data-parallel training of deep neural networks (DNNs). The recent emergence of large language models such as GPT has created the need for new approaches to exploit data-parallelism. Among these, fully-sharded data parallel (FSDP) training is highly popular, yet it still encounters scalability bottlenecks. One reason is that applying compression techniques to FSDP is challenging: as the vast majority of the communication involves the model's weights, direct compression alters convergence and leads to accuracy loss. We present QSDP, a variant of FSDP which supports both gradient and weight quantization with theoretical guarantees, is simple to implement and has essentially no overheads. To derive QSDP we prove that a natural modification of SGD achieves convergence even when we only maintain quantized weights, and thus the domain over which we train consists of quantized points and is, therefore, highly non-convex. We validate this approach by training GPT-family models with up to 1.3 billion parameters on a multi-node cluster. Experiments show that QSDP preserves model accuracy, while completely removing the communication bottlenecks of FSDP, providing end-to-end speedups of up to 2.2x.
****
Efficient Transformed Gaussian Processes for Non-Stationary Dependent Multi-class Classification
Juan Maroñas, Daniel Hernández-Lobato
This work introduces the Efficient Transformed Gaussian Process (ETGP), a new way of creating C stochastic processes characterized by: 1) the C processes are non-stationary, 2) the C processes are dependent by construction without needing a mixing matrix, 3) training and making predictions is very efficient since the number of Gaussian Processes (GP) operations (e.g. inverting the inducing point's covariance matrix) do not depend on the number of processes. This makes the ETGP particularly suited for multi-class problems with a very large number of classes, which are the problems studied in this work. ETGP exploits the recently proposed Transformed Gaussian Process (TGP), a stochastic process specified by transforming a Gaussian Process using an invertible transformation. However, unlike TGP, ETGP is constructed by transforming a single sample from a GP using C invertible transformations. We derive an efficient sparse variational inference algorithm for the proposed model and demonstrate its utility in 5 classification tasks which include low/medium/large datasets and a different number of classes, ranging from just a few to hundreds. Our results show that ETGP, in general, outperforms state-of-the-art methods for multi-class classification based on GPs, and has a lower computational cost (around one order of magnitude smaller).
****
Computational Asymmetries in Robust Classification
Samuele Marro, Michele Lombardi
In the context of adversarial robustness, we make three strongly related contributions. First, we prove that while attacking ReLU classifiers is NP-hard, ensuring their robustness at training time is Σ^2_P-hard (even on a single example). This asymmetry provides a rationale for the fact that robust classifications approaches are frequently fooled in the literature. Second, we show that inference-time robustness certificates are not affected by this asymmetry, by introducing a proof-of-concept approach named Counter-Attack (CA). Indeed, CA displays a reversed asymmetry: running the defense is NP-hard, while attacking it is Σ_2^P-hard. Finally, motivated by our previous result, we argue that adversarial attacks can be used in the context of robustness certification, and provide an empirical evaluation of their effectiveness. As a byproduct of this process, we also release UG100, a benchmark dataset for adversarial attacks.
****
Neural Network Approximations of PDEs Beyond Linearity: A Representational Perspective

Tanya Marwah, Zachary Chase Lipton, Jianfeng Lu, Andrej Risteski

A burgeoning line of research has developed deep neural networks capable of approximating the solutions to high dimensional PDEs, opening related lines of theoretical inquiry focused on explaining how it is that these models appear to evade the curse of dimensionality. However, most theoretical analyses thus far have been limited to linear PDEs. In this work, we take a step towards studying the representational power of neural networks for approximating solutions to nonlinear PDEs. We focus on a class of PDEs known as nonlinear elliptic variational PDEs, whose solutions minimize an Euler-Lagrange energy functional $\mathcal{E}(u) = \int_\Omega L(x, u(x), \nabla u(x)) - f(x) u(x)dx$. We show that if composing a function with Barron norm b with partial derivatives of L produces a function of Barron norm at most $B_L b^p$, the solution to the PDE can be ϵ-approximated in the L^2 sense by a function with Barron norm $O\left(\left(dB_L\right)^{\max\{p \log(1/ \epsilon), p^{\log(1/\epsilon)}\}}\right)$. By a classical result due to Barron (1993), this correspondingly bounds the size of a 2-layer neural network needed to approximate the solution. Treating p, ϵ, B_L as constants, this quantity is polynomial in dimension, thus showing neural networks can evade the curse of dimensionality. Our proof technique involves neurally simulating (preconditioned) gradient in an appropriate Hilbert space, which converges exponentially fast to the solution of the PDE, and such that we can bound the increase of the Barron norm at each iterate. Our results subsume and substantially generalize analogous prior results for linear elliptic PDEs over a unit hypercube.

****

Generative Pretraining for Black-Box Optimization

Satvik Mehul Mashkaria, Siddarth Krishnamoorthy, Aditya Grover

Many problems in science and engineering involve optimizing an expensive black-box function over a high-dimensional space. In the offline model-based optimization (MBO) setting, we assume access to a fixed, offline dataset for pretraining and a small budget for online function evaluations. Prior approaches seek to utilize the offline data to approximate the function or its inverse but are not sufficiently accurate far from the data distribution. We propose BONET, a generative framework for pretraining a novel model-based optimizer using offline datasets. In BONET, we train an autoregressive model on fixed-length trajectories derived from an offline dataset. We design a sampling strategy to synthesize trajectories from offline data using a simple heuristic of rolling out monotonic transitions from low-fidelity to high-fidelity samples. Empirically, we instantiate BONET using a causally masked Transformer (Radford et al., 2019) and evaluate it on Design-Bench (Trabucco et al., 2022), where we rank the best on average, outperforming state-of-the-art baselines.

****

Improved Policy Evaluation for Randomized Trials of Algorithmic Resource Allocation

Aditya Mate, Bryan Wilder, Aparna Taneja, Milind Tambe

We consider the task of evaluating policies of algorithmic resource allocation through randomized controlled trials (RCTs). Such policies are tasked with optimizing the utilization of limited intervention resources, with the goal of maximizing the benefits derived. Evaluation of such allocation policies through RCTs proves difficult, notwithstanding the scale of the trial, because the individuals' outcomes are inextricably interlinked through resource constraints controlling the policy decisions. Our key contribution is to present a new estimator leveraging our proposed novel concept, that involves retrospective reshuffling of participants across experimental arms at the end of an RCT. We identify conditions under which such reassignments are permissible and can be leveraged to construct counterfactual trials, whose outcomes can be accurately ascertained, for free. We prove theoretically that such an estimator is more accurate than common estimators based on sample means – we show that it returns an unbiased estimate and simultaneously reduces variance. We demonstrate the value of our approach through empirical experiments on synthetic, semisynthetic as well as real case study data and show improved estimation accuracy across the board.

****
Multi-Fidelity Covariance Estimation in the Log-Euclidean Geometry
Aimee Maurais, Terrence Alsup, Benjamin Peherstorfer, Youssef Marzouk
We introduce a multi-fidelity estimator of covariance matrices that employs the log-Euclidean geometry of the symmetric positive-definite manifold. The estimator fuses samples from a hierarchy of data sources of differing fidelities and costs for variance reduction while guaranteeing definiteness, in contrast with previous approaches. The new estimator makes covariance estimation tractable in applications where simulation or data collection is expensive; to that end, we develop an optimal sample allocation scheme that minimizes the mean-squared error of the estimator given a fixed budget. Guaranteed definiteness is crucial to metric learning, data assimilation, and other downstream tasks. Evaluations of our approach using data from physical applications (heat conduction, fluid dynamics) demonstrate more accurate metric learning and speedups of more than one order of magnitude compared to benchmarks.
****
Communication-Constrained Bandits under Additive Gaussian Noise
Prathamesh Mayekar, Jonathan Scarlett, Vincent Y. F. Tan
We study a distributed stochastic multi-armed bandit where a client supplies the learner with communication-constrained feedback based on the rewards for the corresponding arm pulls. In our setup, the client must encode the rewards such that the second moment of the encoded rewards is no more than P, and this encoded reward is further corrupted by additive Gaussian noise of variance σ^2; the learner only has access to this corrupted reward. For this setting, we derive an information-theoretic lower bound of $\Omega\left(\sqrt{\frac{KT}{\mathtt{SNR} \wedge 1}} \right)$ on the minimax regret of any scheme, where $\mathtt{SNR}\coloneqq \frac{P}{\sigma^2}$, and K and T are the number of arms and time horizon, respectively. Furthermore, we propose a multi-phase bandit algorithm, $\mathtt{UE}\text{-}\mathtt{UCB}\text{++}$, which matches this lower bound to a minor additive factor. $\mathtt{UE}\text{-}\mathtt{UCB}\text{++}$ performs uniform exploration in its initial phases and then utilizes the upper confidence bound (UCB) bandit algorithm in its final phase. An interesting feature of $\mathtt{UE}\text{-}\mathtt{UCB}\text{++}$ is that the coarser estimates of the mean rewards formed during a uniform exploration phase help to refine the encoding protocol in the next phase, leading to more accurate mean estimates of the rewards in the subsequent phase. This positive reinforcement cycle is critical to reducing the number of uniform exploration rounds and closely matching our lower bound.
****
Nonparametric Density Estimation under Distribution Drift
Alessio Mazzetto, Eli Upfal
We study nonparametric density estimation in non-stationary drift settings. Given a sequence of independent samples taken from a distribution that gradually changes in time, the goal is to compute the best estimate for the current distribution. We prove tight minimax risk bounds for both discrete and continuous smooth densities, where the minimum is over all possible estimates and the maximum is over all possible distributions that satisfy the drift constraints. Our technique handles a broad class of drift models and generalizes previous results on agnostic learning under drift.
****
PAC-Bayesian Generalization Bounds for Adversarial Generative Models
Sokhna Diarra Mbacke, Florence Clerc, Pascal Germain
We extend PAC-Bayesian theory to generative models and develop generalization bounds for models based on the Wasserstein distance and the total variation distance. Our first result on the Wasserstein distance assumes the instance space is bounded, while our second result takes advantage of dimensionality reduction. Our results naturally apply to Wasserstein GANs and Energy-Based GANs, and our bounds provide new training objectives for these two. Although our work is mainly theoretical, we perform numerical experiments showing non-vacuous generalization bounds for Wasserstein GANs on synthetic datasets.
****

Robustness in Multimodal Learning under Train-Test Modality Mismatch
Brandon Mckinzie, Vaishaal Shankar, Joseph Yitan Cheng, Yinfei Yang, Jonathon Shlens, Alexander T Toshev
Multimodal learning is defined as learning over multiple heterogeneous input modalities such as video, audio, and text. In this work, we are concerned with understanding how models behave as the type of modalities differ between training and deployment, a situation that naturally arises in many applications of multimodal learning to hardware platforms. We present a multimodal robustness framework to provide a systematic analysis of common multimodal representation learning methods. Further, we identify robustness short-comings of these approaches and propose two intervention techniques leading to $1.5\times$-$4\times$ robustness improvements on three datasets, AudioSet, Kinetics-400 and ImageNet-Captions. Finally, we demonstrate that these interventions better utilize additional modalities, if present, to achieve competitive results of 44.2 mAP on AudioSet 20K.
****
A Model-free Closeness-of-influence Test for Features in Supervised Learning
Mohammad Mehrabi, Ryan A. Rossi
Understanding the effect of a feature vector $x\in \mathbb{R}^d$ on the response value (label) $y\in \mathbb{R}$ is the cornerstone of many statistical learning problems. Ideally, it is desired to understand how a set of collected features combine together and influence the response value, but this problem is notoriously difficult, due to the high-dimensionality of data and limited number of labeled data points, among many others. In this work, we take a new perspective on this problem, and we study the question of assessing the difference of influence that the two given features have on the response value. We first propose a notion of closeness for the influence of features, and show that our definition recovers the familiar notion of the magnitude of coefficients in the parametric model. We then propose a novel method to test for the closeness of influence in general model-free supervised learning problems. Our proposed test can be used with finite number of samples with control on type I error rate, no matter the ground truth conditional law $\mathcal{L}(Y|X)$. We analyze the power of our test for two general learning problems i) linear regression, and ii) binary classification under mixture of Gaussian models, and show that under the proper choice of score function, an internal component of our test, with sufficient number of samples will achieve full statistical power. We evaluate our findings through extensive numerical simulations, specifically we adopt the datamodel framework (Ilyas, et al., 2022) for CIFAR-10 dataset to identify pairs of training samples with different influence on the trained model via optional black box training mechanisms.
****
Stochastic Gradient Succeeds for Bandits
Jincheng Mei, Zixin Zhong, Bo Dai, Alekh Agarwal, Csaba Szepesvari, Dale Schuurmans
We show that the stochastic gradient bandit algorithm converges to a globally optimal policy at an $O(1/t)$ rate, even with a constant step size. Remarkably, global convergence of the stochastic gradient bandit algorithm has not been previously established, even though it is an old algorithm known to be applicable to bandits. The new result is achieved by establishing two novel technical findings: first, the noise of the stochastic updates in the gradient bandit algorithm satisfies a strong “growth condition” property, where the variance diminishes whenever progress becomes small, implying that additional noise control via diminishing step sizes is unnecessary; second, a form of “weak exploration” is automatically achieved through the stochastic gradient updates, since they prevent the action probabilities from decaying faster than $O(1/t)$, thus ensuring that every action is sampled infinitely often with probability 1. These two findings can be used to show that the stochastic gradient update is already “sufficient” for bandits in the sense that exploration versus exploitation is automatically balanced in a manner that ensures almost sure convergence to a global optimum. These novel theoretical findings are further verified by experimental results.
****
Normalizing Flows for Interventional Density Estimation

Valentyn Melnychuk, Dennis Frauen, Stefan Feuerriegel
Existing machine learning methods for causal inference usually estimate quantities expressed via the mean of potential outcomes (e.g., average treatment effect). However, such quantities do not capture the full information about the distribution of potential outcomes. In this work, we estimate the density of potential outcomes after interventions from observational data. For this, we propose a novel, fully-parametric deep learning method called Interventional Normalizing Flows. Specifically, we combine two normalizing flows, namely (i) a nuisance flow for estimating nuisance parameters and (ii) a target flow for parametric estimation of the density of potential outcomes. We further develop a tractable optimization objective based on a one-step bias correction for efficient and doubly robust estimation of the target flow parameters. As a result, our Interventional Normalizing Flows offer a properly normalized density estimator. Across various experiments, we demonstrate that our Interventional Normalizing Flows are expressive and highly effective, and scale well with both sample size and high-dimensional confounding. To the best of our knowledge, our Interventional Normalizing Flows are the first proper fully-parametric, deep learning method for density estimation of potential outcomes.
****

Reprogramming Pretrained Language Models for Antibody Sequence Infilling
Igor Melnyk, Vijil Chenthamarakshan, Pin-Yu Chen, Payel Das, Amit Dhurandhar, Inkit Padhi, Devleena Das
Antibodies comprise the most versatile class of binding molecules, with numerous applications in biomedicine. Computational design of antibodies involves generating novel and diverse sequences, while maintaining structural consistency. Unique to antibodies, designing the complementarity-determining region (CDR), which determines the antigen binding affinity and specificity, creates its own unique challenges. Recent deep learning models have shown impressive results, however the limited number of known antibody sequence/structure pairs frequently leads to degraded performance, particularly lacking diversity in the generated sequences. In our work we address this challenge by leveraging Model Reprogramming (MR), which repurposes pretrained models on a source language to adapt to the tasks that are in a different language and have scarce data - where it may be difficult to train a high-performing model from scratch or effectively fine-tune an existing pre-trained model on the specific task. Specifically, we introduce ReprogBert in which a pretrained English language model is repurposed for protein sequence infilling - thus considers cross-language adaptation using less data. Results on antibody design benchmarks show that our model on low-resourced antibody sequence dataset provides highly diverse CDR sequences, up to more than a two-fold increase of diversity over the baselines, without losing structural integrity and naturalness. The generated sequences also demonstrate enhanced antigen binding specificity and virus neutralization ability. Code is available at https://github.com/IBM/ReprogBERT
****

Superhuman Fairness
Omid Memarrast, Linh Vu, Brian D Ziebart
The fairness of machine learning-based decisions has become an increasingly important focus in the design of supervised machine learning methods. Most fairness approaches optimize a specified trade-off between performance measure(s) (e.g., accuracy, log loss, or AUC) and fairness metric(s) (e.g., demographic parity, equalized odds). This begs the question: are the right performance-fairness trade-offs being specified? We instead re-cast fair machine learning as an imitation learning task by introducing superhuman fairness, which seeks to simultaneously outperform human decisions on multiple predictive performance and fairness measures. We demonstrate the benefits of this approach given suboptimal decisions.
****

A Model-Based Method for Minimizing CVaR and Beyond
Si Yi Meng, Robert M. Gower
We develop a variant of the stochastic prox-linear method for minimizing the Conditional Value-at-Risk (CVaR) objective. CVaR is a risk measure focused on minim

izing worst-case performance, defined as the average of the top quantile of the losses. In machine learning, such a risk measure is useful to train more robust models. Although the stochastic subgradient method (SGM) is a natural choice for minimizing the CVaR objective, we show that our stochastic prox-linear (SPL+) algorithm can better exploit the structure of the objective, while still providing a convenient closed form update. Our SPL+ method also adapts to the scaling of the loss function, which allows for easier tuning. We then specialize a general convergence theorem for SPL+ to our setting, and show that it allows for a wider selection of step sizes compared to SGM. We support this theoretical finding experimentally.

****

Tuning Language Models as Training Data Generators for Augmentation-Enhanced Few-Shot Learning

Yu Meng, Martin Michalski, Jiaxin Huang, Yu Zhang, Tarek Abdelzaher, Jiawei Han

Recent studies have revealed the intriguing few-shot learning ability of pretrained language models (PLMs): They can quickly adapt to a new task when fine-tuned on a small amount of labeled data formulated as prompts, without requiring abundant task-specific annotations. Despite their promising performance, most existing few-shot approaches that only learn from the small training set still underperform fully supervised training by nontrivial margins. In this work, we study few-shot learning with PLMs from a different perspective: We first tune an autoregressive PLM on the few-shot samples and then use it as a generator to synthesize a large amount of novel training samples which augment the original training set. To encourage the generator to produce label-discriminative samples, we train it via weighted maximum likelihood where the weight of each token is automatically adjusted based on a discriminative meta-learning objective. A classification PLM can then be fine-tuned on both the few-shot and the synthetic samples with regularization for better generalization and stability. Our approach FewGen achieves an overall better result across seven classification tasks of the GLUE benchmark than existing few-shot learning methods, improving no-augmentation methods by 5+ average points, and outperforming augmentation methods by 3+ average points.

****

On Preemption and Learning in Stochastic Scheduling

Nadav Merlis, Hugo Richard, Flore Sentenac, Corentin Odic, Mathieu Molina, Vianney Perchet

We study single-machine scheduling of jobs, each belonging to a job type that determines its duration distribution. We start by analyzing the scenario where the type characteristics are known and then move to two learning scenarios where the types are unknown: non-preemptive problems, where each started job must be completed before moving to another job; and preemptive problems, where job execution can be paused in the favor of moving to a different job. In both cases, we design algorithms that achieve sublinear excess cost, compared to the performance with known types, and prove lower bounds for the non-preemptive case. Notably, we demonstrate, both theoretically and through simulations, how preemptive algorithms can greatly outperform non-preemptive ones when the durations of different job types are far from one another, a phenomenon that does not occur when the type durations are known.

****

Quantile Credit Assignment

Thomas Mesnard, Wenqi Chen, Alaa Saade, Yunhao Tang, Mark Rowland, Theophane Weber, Clare Lyle, Audrunas Gruslys, Michal Valko, Will Dabney, Georg Ostrovski, Eric Moulines, Remi Munos

In reinforcement learning, the credit assignment problem is to distinguish luck from skill, that is, separate the inherent randomness in the environment from the controllable effects of the agent’s actions. This paper proposes two novel algorithms, Quantile Credit Assignment (QCA) and Hindsight QCA (HQCA), which incorporate distributional value estimation to perform credit assignment. QCA uses a network that predicts the quantiles of the return distribution, whereas HQCA additionally incorporates information about the future. Both QCA and HQCA have the a

ppealing interpretation of leveraging an estimate of the quantile level of the r
eturn (interpreted as the level of "luck") in order to derive a "luck-dependent"
 baseline for policy gradient methods. We show theoretically that this approach
gives an unbiased policy gradient estimate that can yield significant variance r
eductions over a standard value estimate baseline. QCA and HQCA significantly ou
tperform prior state-of-the-art methods on a range of extremely difficult credit
 assignment problems.
****

Is Consensus Acceleration Possible in Decentralized Optimization over Slowly Tim
e-Varying Networks?
Dmitry Metelev, Alexander Rogozin, Dmitry Kovalev, Alexander Gasnikov
We consider decentralized optimization problems where one aims to minimize a sum
 of convex smooth objective functions distributed between nodes in the network.
The links in the network can change from time to time. For the setting when the
amount of changes is arbitrary, lower complexity bounds and corresponding optima
l algorithms are known, and the consensus acceleration is not possible. However,
 in practice the magnitude of network changes may be limited. We derive lower co
mplexity bounds for several regimes of velocity of networks changes. Moreover, w
e show how to obtain accelerated communication rates for a certain class of time
-varying graphs using a specific consensus algorithm.
****

Towards Theoretical Understanding of Inverse Reinforcement Learning
Alberto Maria Metelli, Filippo Lazzati, Marcello Restelli
Inverse reinforcement learning (IRL) denotes a powerful family of algorithms for
 recovering a reward function justifying the behavior demonstrated by an expert
agent. A well-known limitation of IRL is the ambiguity in the choice of the rewa
rd function, due to the existence of multiple rewards that explain the observed
behavior. This limitation has been recently circumvented by formulating IRL as t
he problem of estimating the feasible reward set, i.e., the region of the reward
s compatible with the expert’s behavior. In this paper, we make a step towards c
losing the theory gap of IRL in the case of finite-horizon problems with a gener
ative model. We start by formally introducing the problem of estimating the feas
ible reward set, the corresponding PAC requirement, and discussing the propertie
s of particular classes of rewards. Then, we provide the first minimax lower bou
nd on the sample complexity for the problem of estimating the feasible reward se
t of order ${\Omega}\left(\frac{H^3SA}{\epsilon^2} \left(\log \left(\frac{1}{\
delta}\right) + S \right)\right)$, being S and A the number of states and ac
tions respectively, H the horizon, ϵ the desired accuracy, and $\delt
a$ the confidence. We analyze the sample complexity of a uniform sampling strate
gy (US-IRL), proving a matching upper bound up to logarithmic factors. Finally,
we outline several open questions in IRL and propose future research directions.
****

Quantum Policy Gradient Algorithm with Optimized Action Decoding
Nico Meyer, Daniel Scherer, Axel Plinge, Christopher Mutschler, Michael Hartmann
Quantum machine learning implemented by variational quantum circuits (VQCs) is c
onsidered a promising concept for the noisy intermediate-scale quantum computing
 era. Focusing on applications in quantum reinforcement learning, we propose an
action decoding procedure for a quantum policy gradient approach. We introduce a
 quality measure that enables us to optimize the classical post-processing requi
red for action selection, inspired by local and global quantum measurements. The
 resulting algorithm demonstrates a significant performance improvement in sever
al benchmark environments. With this technique, we successfully execute a full t
raining routine on a 5-qubit hardware device. Our method introduces only negligi
ble classical overhead and has the potential to improve VQC-based algorithms bey
ond the field of quantum reinforcement learning.
****

Training Deep Surrogate Models with Large Scale Online Learning
Lucas Thibaut Meyer, Marc Schouler, Robert Alexander Caulk, Alejandro Ribes, Bru
no Raffin
The spatiotemporal resolution of Partial Differential Equations (PDEs) plays imp

ortant roles in the mathematical description of the world’s physical phenomena. In general, scientists and engineers solve PDEs numerically by the use of computationally demanding solvers. Recently, deep learning algorithms have emerged as a viable alternative for obtaining fast solutions for PDEs. Models are usually trained on synthetic data generated by solvers, stored on disk and read back for training. This paper advocates that relying on a traditional static dataset to train these models does not allow the full benefit of the solver to be used as a data generator. It proposes an open source online training framework for deep surrogate models. The framework implements several levels of parallelism focused on simultaneously generating numerical simulations and training deep neural networks. This approach suppresses the I/O and storage bottleneck associated with disk-loaded datasets, and opens the way to training on significantly larger datasets. Experiments compare the offline and online training of four surrogate models, including state-of-the-art architectures. Results indicate that exposing deep surrogate models to more dataset diversity, up to hundreds of GB, can increase model generalization capabilities. Fully connected neural networks, Fourier Neural Operator (FNO), and Message Passing PDE Solver prediction accuracy is improved by 68%, 16% and 7%, respectively.

****

MANSA: Learning Fast and Slow in Multi-Agent Systems

David Henry Mguni, Haojun Chen, Taher Jafferjee, Jianhong Wang, Longfei Yue, Xidong Feng, Stephen Marcus Mcaleer, Feifei Tong, Jun Wang, Yaodong Yang

In multi-agent reinforcement learning (MARL), independent learning (IL) often shows remarkable performance and easily scales with the number of agents. Yet, using IL can be inefficient and runs the risk of failing to successfully train, particularly in scenarios that require agents to coordinate their actions. Using centralised learning (CL) enables MARL agents to quickly learn how to coordinate their behaviour but employing CL everywhere is often prohibitively expensive in real-world applications. Besides, using CL in value-based methods often needs strong representational constraints (e.g. individual-global-max condition) that can lead to poor performance if violated. In this paper, we introduce a novel plug & play IL framework named Multi-Agent Network Selection Algorithm (MANSA) which selectively employs CL only at states that require coordination. At its core, MANSA has an additional agent that uses switching controls to quickly learn the best states to activate CL during training, using CL only where necessary and vastly reducing the computational burden of CL. Our theory proves MANSA preserves cooperative MARL convergence properties, boosts IL performance and can optimally make use of a fixed budget on the number CL calls. We show empirically in Level-based Foraging (LBF) and StarCraft Multi-agent Challenge (SMAC) that MANSA achieves fast, superior and more reliable performance while making 40% fewer CL calls in SMAC and using CL at only 1% CL calls in LBF.

****

Representation Learning with Multi-Step Inverse Kinematics: An Efficient and Optimal Approach to Rich-Observation RL

Zakaria Mhammedi, Dylan J Foster, Alexander Rakhlin

We study the design of sample-efficient algorithms for reinforcement learning in the presence of rich, high-dimensional observations, formalized via the Block MDP problem. Existing algorithms suffer from either 1) computational intractability, 2) strong statistical assumptions that are not necessarily satisfied in practice, or 3) suboptimal sample complexity. We address these issues by providing the first computationally efficient algorithm that attains rate-optimal sample complexity with respect to the desired accuracy level, with minimal statistical assumptions. Our algorithm, MusIK, combines exploration with representation learning based on multi-step inverse kinematics, a learning objective in which the aim is to predict the current action from the current observation and observations in the (potentially distant) future. MusIK is simple and flexible, and can efficiently take advantage of general-purpose function approximation. Our analysis of MusIK leverages several new techniques tailored to non-optimistic algorithms for reward-free exploration, which we anticipate will find broader use.

****

Single Point-Based Distributed Zeroth-Order Optimization with a Non-Convex Stochastic Objective Function
Elissa Mhanna, Mohamad Assaad
Zero-order (ZO) optimization is a powerful tool for dealing with realistic constraints. On the other hand, the gradient-tracking (GT) technique proved to be an efficient method for distributed optimization aiming to achieve consensus. However, it is a first-order (FO) method that requires knowledge of the gradient, which is not always possible in practice. In this work, we introduce a zero-order distributed optimization method based on a one-point estimate of the gradient tracking technique. We prove that this new technique converges with a single noisy function query at a time in the non-convex setting. We then establish a convergence rate of $O(\frac{1}{\sqrt[3]{K}})$ after a number of iterations K, which competes with that of $O(\frac{1}{\sqrt[4]{K}})$ of its centralized counterparts. Finally, a numerical example validates our theoretical results.
****
Learning Instance-Specific Augmentations by Capturing Local Invariances
Ning Miao, Tom Rainforth, Emile Mathieu, Yann Dubois, Yee Whye Teh, Adam Foster, Hyunjik Kim
We introduce InstaAug, a method for automatically learning input-specific augmentations from data. Previous methods for learning augmentations have typically assumed independence between the original input and the transformation applied to that input. This can be highly restrictive, as the invariances we hope our augmentation will capture are themselves often highly input dependent. InstaAug instead introduces a learnable invariance module that maps from inputs to tailored transformation parameters, allowing local invariances to be captured. This can be simultaneously trained alongside the downstream model in a fully end-to-end manner, or separately learned for a pre-trained model. We empirically demonstrate that InstaAug learns meaningful input-dependent augmentations for a wide range of transformation classes, which in turn provides better performance on both supervised and self-supervised tasks.
****
Path Neural Networks: Expressive and Accurate Graph Neural Networks
Gaspard Michel, Giannis Nikolentzos, Johannes F. Lutzeyer, Michalis Vazirgiannis
Graph neural networks (GNNs) have recently become the standard approach for learning with graph-structured data. Prior work has shed light into their potential, but also their limitations. Unfortunately, it was shown that standard GNNs are limited in their expressive power. These models are no more powerful than the 1-dimensional Weisfeiler-Leman (1-WL) algorithm in terms of distinguishing non-isomorphic graphs. In this paper, we propose Path Neural Networks (PathNNs), a model that updates node representations by aggregating paths emanating from nodes. We derive three different variants of the PathNN model that aggregate single shortest paths, all shortest paths and all simple paths of length up to K. We prove that two of these variants are strictly more powerful than the 1-WL algorithm, and we experimentally validate our theoretical results. We find that PathNNs can distinguish pairs of non-isomorphic graphs that are indistinguishable by 1-WL, while our most expressive PathNN variant can even distinguish between 3-WL indistinguishable graphs. The different PathNN variants are also evaluated on graph classification and graph regression datasets, where in most cases, they outperform the baseline methods.
****
Learning to acquire novel cognitive tasks with evolution, plasticity and meta-meta-learning
Thomas Miconi
A hallmark of intelligence is the ability to autonomously learn new flexible, cognitive behaviors - that is, behaviors where the appropriate action depends not just on immediate stimuli (as in simple reflexive stimulus-response associations), but on contextual information that must be adequately acquired, stored and processed. While many meta-learning algorithms can design agents that autonomously learn new tasks, cognitive tasks adds another level of learning and memory to typical “learning-to-learn” problems. Here we evolve neural networks, endowed wit

h plastic connections and neuromodulation, over a sizable set of simple cognitiv e tasks adapted from a computational neuroscience framework. The resulting evolv ed networks can automatically modify their own connectivity to acquire a novel s imple cognitive task, never seen during evolution, from stimuli and rewards alon e, through the spontaneous operation of their evolved neural organization and pl asticity system. Our results emphasize the importance of carefully considering t he multiple learning loops involved in the emergence of intelligent behavior.
****
Generative Decoding of Visual Stimuli
Eleni Miliotou, Panagiotis Kyriakis, Jason D Hinman, Andrei Irimia, Paul Bogdan
Reconstructing natural images from fMRI recordings is a challenging task of grea t importance in neuroscience. The current architectures are bottlenecked because they fail to effectively capture the hierarchical processing of visual stimuli that takes place in the human brain. Motivated by that fact, we introduce a nove l neural network architecture for the problem of neural decoding. Our architectu re uses Hierarchical Variational Autoencoders (HVAEs) to learn meaningful repres entations of natural images and leverages their latent space hierarchy to learn voxel-to-image mappings. By mapping the early stages of the visual pathway to th e first set of latent variables and the higher visual cortex areas to the deeper layers in the latent hierarchy, we are able to construct a latent variable neur al decoding model that replicates the hierarchical visual information processing . Our model achieves better reconstructions compared to the state of the art and our ablation study indicates that the hierarchical structure of the latent spac e is responsible for that performance.
****
Cooperative Multi-Agent Reinforcement Learning: Asynchronous Communication and L inear Function Approximation
Yifei Min, Jiafan He, Tianhao Wang, Quanquan Gu
We study multi-agent reinforcement learning in the setting of episodic Markov de cision processes, where many agents cooperate via communication through a centra l server. We propose a provably efficient algorithm based on value iteration tha t can simultaneously allow asynchronous communication and guarantee the benefit of cooperation with low communication complexity. Under linear function approxim ation, we prove that our algorithm enjoys a $\tilde{\mathcal{O}}(d^{3/2}H^2\sqrt {K})$ regret upper bound with $\tilde{\mathcal{O}}(dHM^2)$ communication complex ity, where d is the feature dimension, H is the horizon length, M is the t otal number of agents, and K is the total number of episodes. We also provide a lower bound showing that an $\Omega(dM)$ communication complexity is necessary to improve the performance through collaboration.
****
Directed Chain Generative Adversarial Networks
Ming Min, Ruimeng Hu, Tomoyuki Ichiba
Real-world data can be multimodal distributed, e.g., data describing the opinion divergence in a community, the interspike interval distribution of neurons, and the oscillators natural frequencies. Generating multimodal distributed real-wor ld data has become a challenge to existing generative adversarial networks (GANs). For example, it is often observed that Neural SDEs have only demonstrated suc cessfully performance mainly in generating unimodal time series datasets. In thi s paper, we propose a novel time series generator, named directed chain GANs (DC -GANs), which inserts a time series dataset (called a neighborhood process of th e directed chain or input) into the drift and diffusion coefficients of the dire cted chain SDEs with distributional constraints. DC-GANs can generate new time s eries of the same distribution as the neighborhood process, and the neighborhood process will provide the key step in learning and generating multimodal distrib uted time series. The proposed DC-GANs are examined on four datasets, including two stochastic models from social sciences and computational neuroscience, and t wo real-world datasets on stock prices and energy consumption. To our best knowl edge, DC-GANs are the first work that can generate multimodal time series data a nd consistently outperforms state-of-the-art benchmarks with respect to measures of distribution, data similarity, and predictive ability.

****
An Information-Theoretic Analysis of Nonstationary Bandit Learning
Seungki Min, Daniel Russo
In nonstationary bandit learning problems, the decision-maker must continually gather information and adapt their action selection as the latent state of the environment evolves. In each time period, some latent optimal action maximizes expected reward under the environment state. We view the optimal action sequence as a stochastic process, and take an information-theoretic approach to analyze attainable performance. We bound per-period regret in terms of the entropy rate of the optimal action process. The bound applies to a wide array of problems studied in the literature and reflects the problem's information structure through its information-ratio.
****
On the Convergence of Gradient Flow on Multi-layer Linear Models
Hancheng Min, Rene Vidal, Enrique Mallada
In this paper, we analyze the convergence of gradient flow on a multi-layer linear model with a loss function of the form $f(W_1W_2\cdots W_L)$. We show that when f satisfies the gradient dominance property, proper weight initialization leads to exponential convergence of the gradient flow to a global minimum of the loss. Moreover, the convergence rate depends on two trajectory-specific quantities that are controlled by the weight initialization: the imbalance matrices, which measure the difference between the weights of adjacent layers, and the least singular value of the weight product $W=W_1W_2\cdots W_L$. Our analysis exploits the fact that the gradient of the overparameterized loss can be written as the composition of the non-overparametrized gradient with a time-varying (weight-dependent) linear operator whose smallest eigenvalue controls the convergence rate. The key challenge we address is to derive a uniform lower bound for this time-varying eigenvalue that lead to improved rates for several multi-layer network models studied in the literature.
****
Optimal Sets and Solution Paths of ReLU Networks
Aaron Mishkin, Mert Pilanci
We develop an analytical framework to characterize the set of optimal ReLU neural networks by reformulating the non-convex training problem as a convex program. We show that the global optima of the convex parameterization are given by a polyhedral set and then extend this characterization to the optimal set of the non-convex training objective. Since all stationary points of the ReLU training problem can be represented as optima of sub-sampled convex programs, our work provide a general expression for all critical points of the non-convex objective. We then leverage our results to provide an optimal pruning algorithm for computing minimal networks, establish conditions for the regularization path of ReLU networks to be continuous, and develop sensitivity results for minimal ReLU networks.
****
The Numerical Stability of Hyperbolic Representation Learning
Gal Mishne, Zhengchao Wan, Yusu Wang, Sheng Yang
The hyperbolic space is widely used for representing hierarchical datasets due to its ability to embed trees with small distortion. However, this property comes at a price of numerical instability such that training hyperbolic learning models will sometimes lead to catastrophic NaN problems, encountering unrepresentable values in floating point arithmetic. In this work, we analyze the limitations of two popular models for the hyperbolic space, namely, the Poincaré ball and the Lorentz model. We find that, under the 64-bit arithmetic system, the Poincaré ball has a relatively larger capacity than the Lorentz model for correctly representing points. However, the Lorentz model is superior to the Poincaré ball from the perspective of optimization, which we theoretically validate. To address these limitations, we identify one Euclidean parametrization of the hyperbolic space which can alleviate these issues. We further extend this Euclidean parametrization to hyperbolic hyperplanes and demonstrate its effectiveness in improving the performance of hyperbolic SVM.
****

DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvatur
e
Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, Chelsea F
inn
The increasing fluency and widespread usage of large language models (LLMs) high
light the desirability of corresponding tools aiding detection of LLM-generated
text. In this paper, we identify a property of the structure of an LLM’s probabi
lity function that is useful for such detection. Specifically, we demonstrate th
at text sampled from an LLM tends to occupy negative curvature regions of the mo
del’s log probability function. Leveraging this observation, we then define a ne
w curvature-based criterion for judging if a passage is generated from a given L
LM. This approach, which we call DetectGPT, does not require training a separate
 classifier, collecting a dataset of real or generated passages, or explicitly w
atermarking generated text. It uses only log probabilities computed by the model
 of interest and random perturbations of the passage from another generic pre-tr
ained language model (e.g., T5). We find DetectGPT is more discriminative than e
xisting zero-shot methods for model sample detection, notably improving detectio
n of fake news articles generated by 20B parameter GPT-NeoX from 0.81 AUROC for
the strongest zero-shot baseline to 0.95 AUROC for DetectGPT.
****

Diffusion Based Representation Learning
Sarthak Mittal, Korbinian Abstreiter, Stefan Bauer, Bernhard Schölkopf, Arash Me
hrjou
Diffusion-based methods, represented as stochastic differential equations on a c
ontinuous-time domain, have recently proven successful as non-adversarial genera
tive models. Training such models relies on denoising score matching, which can
be seen as multi-scale denoising autoencoders. Here, we augment the denoising sc
ore matching framework to enable representation learning without any supervised
signal. GANs and VAEs learn representations by directly transforming latent code
s to data samples. In contrast, the introduced diffusion-based representation le
arning relies on a new formulation of the denoising score matching objective and
 thus encodes the information needed for denoising. We illustrate how this diffe
rence allows for manual control of the level of details encoded in the represent
ation. Using the same approach, we propose to learn an infinite-dimensional late
nt code that achieves improvements on state-of-the-art models on semi-supervised
 image classification. We also compare the quality of learned representations of
 diffusion score matching with other methods like autoencoder and contrastively
trained systems through their performances on downstream tasks. Finally, we also
 ablate with a different SDE formulation for diffusion models and show that the
benefits on downstream tasks are still present on changing the underlying differ
ential equation.
****

Disentangled Multiplex Graph Representation Learning
Yujie Mo, Yajie Lei, Jialie Shen, Xiaoshuang Shi, Heng Tao Shen, Xiaofeng Zhu
Unsupervised multiplex graph representation learning (UMGRL) has received increa
sing interest, but few works simultaneously focused on the common and private in
formation extraction. In this paper, we argue that it is essential for conductin
g effective and robust UMGRL to extract complete and clean common information, a
s well as more-complementarity and less-noise private information. To achieve th
is, we first investigate disentangled representation learning for the multiplex
graph to capture complete and clean common information, as well as design a cont
rastive constraint to preserve the complementarity and remove the noise in the p
rivate information. Moreover, we theoretically analyze that the common and priva
te representations learned by our method are provably disentangled and contain m
ore task-relevant and less task-irrelevant information to benefit downstream tas
ks. Extensive experiments verify the superiority of the proposed method in terms
 of different downstream tasks.
****

A Unified Audio-Visual Learning Framework for Localization, Separation, and Reco
gnition

Shentong Mo, Pedro Morgado
The ability to accurately recognize, localize and separate sound sources is fundamental to any audio-visual perception task. Historically, these abilities were tackled separately, with several methods developed independently for each task. However, given the interconnected nature of source localization, separation, and recognition, independent models are likely to yield suboptimal performance as they fail to capture the interdependence between these tasks. To address this problem, we propose a unified audio-visual learning framework (dubbed OneAVM) that integrates audio and visual cues for joint localization, separation, and recognition. OneAVM comprises a shared audio-visual encoder and task-specific decoders trained with three objectives. The first objective aligns audio and visual representations through a localized audio-visual correspondence loss. The second tackles visual source separation using a traditional mix-and-separate framework. Finally, the third objective reinforces visual feature separation and localization by mixing images in pixel space and aligning their representations with those of all corresponding sound sources. Extensive experiments on MUSIC, VGG-Instruments, VGG-Music, and VGGSound datasets demonstrate the effectiveness of OneAVM for all three tasks, audio-visual source localization, separation, and nearest neighbor recognition, and empirically demonstrate a strong positive transfer between them.
****

Pruning via Sparsity-indexed ODE: a Continuous Sparsity Viewpoint
Zhanfeng Mo, Haosen Shi, Sinno Jialin Pan
Neural pruning, which involves identifying the optimal sparse subnetwork, is a key technique for reducing the complexity and improving the efficiency of deep neural networks. To address the challenge of solving neural pruning at a specific sparsity level directly, we investigate the evolution of optimal subnetworks with continuously increasing sparsity, which can provide insight into how to transform an unpruned dense model into an optimal subnetwork with any desired level of sparsity. In this paper, we proposed a novel pruning framework, coined Sparsity-indexed ODE (SpODE) that provides explicit guidance on how to best preserve model performance while ensuring an infinitesimal increase in model sparsity. On top of this, we develop a pruning algorithm, termed Pruning via Sparsity-indexed ODE (PSO), that enables effective pruning via traveling along the SpODE path. Empirical experiments show that PSO achieves either better or comparable performance compared to state-of-the-art baselines across various pruning settings.
****

Text-To-Concept (and Back) via Cross-Model Alignment
Mazda Moayeri, Keivan Rezaei, Maziar Sanjabi, Soheil Feizi
We observe that the mapping between an image’s representation in one model to its representation in another can be learned surprisingly well with just a linear layer, even across diverse models. Building on this observation, we propose text-to-concept, where features from a fixed pretrained model are aligned linearly to the CLIP space, so that text embeddings from CLIP’s text encoder become directly comparable to the aligned features. With text-to-concept, we convert fixed off-the-shelf vision encoders to surprisingly strong zero-shot classifiers for free, with accuracy at times even surpassing that of CLIP, despite being much smaller models and trained on a small fraction of the data compared to CLIP. We show other immediate use-cases of text-to-concept, like building concept bottleneck models with no concept supervision, diagnosing distribution shifts in terms of human concepts, and retrieving images satisfying a set of text-based constraints. Lastly, we demonstrate the feasibility of concept-to-text, where vectors in a model’s feature space are decoded by first aligning to the CLIP before being fed to a GPT-based generative model. Our work suggests existing deep models, with presumably diverse architectures and training, represent input samples relatively similarly, and a two-way communication across model representation spaces and to humans (through language) is viable.
****

A Fast, Well-Founded Approximation to the Empirical Neural Tangent Kernel
Mohamad Amin Mohamadi, Wonho Bae, Danica J. Sutherland

Empirical neural tangent kernels (eNTKs) can provide a good understanding of a g
iven network’s representation: they are often far less expensive to compute and
applicable more broadly than infinite-width NTKs. For networks with O output u
nits (e.g. an O-class classifier), however, the eNTK on N inputs is of size
$NO \times NO$, taking $\mathcal O\big((N O)^2\big)$ memory and up to $\mathcal
 O\big((N O)^3 \big)$ computation to use. Most existing applications have there
fore used one of a handful of approximations yielding $N \times N$ kernel matric
es, saving orders of magnitude of computation, but with limited to no justificat
ion. We prove that one such approximation, which we call "sum of logits," conver
ges to the true eNTK at initialization. Our experiments demonstrate the quality
of this approximation for various uses across a range of settings.
****

Special Properties of Gradient Descent with Large Learning Rates
Amirkeivan Mohtashami, Martin Jaggi, Sebastian U Stich
When training neural networks, it has been widely observed that a large step siz
e is essential in stochastic gradient descent (SGD) for obtaining superior model
s. However, the effect of large step sizes on the success of SGD is not well und
erstood theoretically. Several previous works have attributed this success to th
e stochastic noise present in SGD. However, we show through a novel set of exper
iments that the stochastic noise is not sufficient to explain good non-convex tr
aining, and that instead the effect of a large learning rate itself is essential
 for obtaining best performance.We demonstrate the same effects also in the nois
e-less case, i.e. for full-batch GD. We formally prove that GD with large step s
ize —on certain non-convex function classes — follows a different trajectory tha
n GD with a small step size, which can lead to convergence to a global minimum i
nstead of a local one. Our settings provide a framework for future analysis whic
h allows comparing algorithms based on behaviors that can not be observed in the
 traditional settings.
****

Neural Inverse Operators for Solving PDE Inverse Problems
Roberto Molinaro, Yunan Yang, Björn Engquist, Siddhartha Mishra
A large class of inverse problems for PDEs are only well-defined as mappings fro
m operators to functions. Existing operator learning frameworks map functions to
 functions and need to be modified to learn inverse maps from data. We propose a
 novel architecture termed Neural Inverse Operators (NIOs) to solve these PDE in
verse problems. Motivated by the underlying mathematical structure, NIO is based
 on a suitable composition of DeepONets and FNOs to approximate mappings from op
erators to functions. A variety of experiments are presented to demonstrate that
 NIOs significantly outperform baselines and solve PDE inverse problems robustly
, accurately and are several orders of magnitude faster than existing direct and
 PDE-constrained optimization methods.
****

Input uncertainty propagation through trained neural networks
Paul Monchot, Loic Coquelin, Sébastien Julien Petit, Sébastien Marmin, Erwan Le
Pennec, Nicolas Fischer
When physical sensors are involved, such as image sensors, the uncertainty over
the input data is often a major component of the output uncertainty of machine l
earning models. In this work, we address the problem of input uncertainty propag
ation through trained neural networks. We do not rely on a Gaussian distribution
 assumption of the output or of any intermediate layer. We propagate instead a G
aussian Mixture Model (GMM) that offers much more flexibility, using the Split&M
erge algorithm. This paper’s main contribution is the computation of a Wasserste
in criterion to control the Gaussian splitting procedure for which theoretical g
uarantees of convergence on the output distribution estimates are derived. The m
ethodology is tested against a wide range of datasets and networks. It shows rob
ustness, and genericity and offers highly accurate output probability density fu
nction estimation while maintaining a reasonable computational cost compared wit
h the standard Monte Carlo (MC) approach.
****

Compressing Tabular Data via Latent Variable Estimation

Andrea Montanari, Eric Weiner
Data used for analytics and machine learning often take the form of tables with categorical entries. We introduce a family of lossless compression algorithms for such data that proceed in four steps: (i) Estimate latent variables associated to rows and columns; (ii) Partition the table in blocks according to the row/column latents; (iii) Apply a sequential (e.g. Lempel-Ziv) coder to each of the blocks; (iv) Append a compressed encoding of the latents. We evaluate this approach on several benchmark datasets, and study optimal compression in a probabilistic model for tabular data, whereby latent values are independent and table entries are conditionally independent given the latent values. We prove that the model has a well defined entropy rate and satisfies an asymptotic equipartition property. We also prove that classical compression schemes such as Lempel-Ziv and finite-state encoders do not achieve this rate. On the other hand, the latent estimation strategy outlined above achieves the optimal rate.
****

An SDE for Modeling SAM: Theory and Insights
Enea Monzio Compagnoni, Luca Biggio, Antonio Orvieto, Frank Norbert Proske, Hans Kersting, Aurelien Lucchi
We study the SAM (Sharpness-Aware Minimization) optimizer which has recently attracted a lot of interest due to its increased performance over more classical variants of stochastic gradient descent. Our main contribution is the derivation of continuous-time models (in the form of SDEs) for SAM and two of its variants, both for the full-batch and mini-batch settings. We demonstrate that these SDEs are rigorous approximations of the real discrete-time algorithms (in a weak sense, scaling linearly with the learning rate). Using these models, we then offer an explanation of why SAM prefers flat minima over sharp ones – by showing that it minimizes an implicitly regularized loss with a Hessian-dependent noise structure. Finally, we prove that SAM is attracted to saddle points under some realistic conditions. Our theoretical results are supported by detailed experiments.
****

Learning Deductive Reasoning from Synthetic Corpus based on Formal Logic
Terufumi Morishita, Gaku Morio, Atsuki Yamaguchi, Yasuhiro Sogawa
We study a synthetic corpus based approach for language models (LMs) to acquire logical deductive reasoning ability. The previous studies generated deduction examples using specific sets of deduction rules. However, these rules were limited or otherwise arbitrary. This can limit the generalizability of acquired deductive reasoning ability. We rethink this and adopt a well-grounded set of deduction rules based on formal logic theory, which can derive any other deduction rules when combined in a multistep way. We empirically verify that LMs trained on the proposed corpora, which we name $\textbf{FLD}$ ($\textbf{F}$ormal $\textbf{L}$ogic $\textbf{D}$eduction), acquire more generalizable deductive reasoning ability. Furthermore, we identify the aspects of deductive reasoning ability on which deduction corpora can enhance LMs and those on which they cannot. Finally, on the basis of these results, we discuss the future directions for applying deduction corpora or other approaches for each aspect. We release the code, data, and models.
****

WL meet VC
Christopher Morris, Floris Geerts, Jan Tönshoff, Martin Grohe
Recently, many works studied the expressive power of graph neural networks (GNNs) by linking it to the 1-dimensional Weisfeiler-Leman algorithm ($1\text{-}\mathsf{WL}$). Here, the $1\text{-}\mathsf{WL}$ is a well-studied heuristic for the graph isomorphism problem, which iteratively colors or partitions a graph’s vertex set. While this connection has led to significant advances in understanding and enhancing GNNs’ expressive power, it does not provide insights into their generalization performance, i.e., their ability to make meaningful predictions beyond the training set. In this paper, we study GNNs’ generalization ability through the lens of Vapnik-Chervonenkis (VC) dimension theory in two settings, focusing on graph-level predictions. First, when no upper bound on the graphs’ order is known, we show that the bitlength of GNNs’ weights tightly bounds their VC dim

ension. Further, we derive an upper bound for GNNs' VC dimension using the numbe r of colors produced by the $1\text{-}\mathsf{WL}$. Secondly, when an upper boun d on the graphs' order is known, we show a tight connection between the number o f graphs distinguishable by the $1\text{-}\mathsf{WL}$ and GNNs' VC dimension. O ur empirical study confirms the validity of our theoretical findings.
****

ReLOAD: Reinforcement Learning with Optimistic Ascent-Descent for Last-Iterate C onvergence in Constrained MDPs
Ted Moskovitz, Brendan O'Donoghue, Vivek Veeriah, Sebastian Flennerhag, Satinder Singh, Tom Zahavy
In recent years, reinforcement learning (RL) has been applied to real-world prob lems with increasing success. Such applications often require to put constraints on the agent's behavior. Existing algorithms for constrained RL (CRL) rely on g radient descent-ascent, but this approach comes with a caveat. While these algor ithms are guaranteed to converge on average, they do not guarantee last-iterate convergence, i.e., the current policy of the agent may never converge to the opt imal solution. In practice, it is often observed that the policy alternates betw een satisfying the constraints and maximizing the reward, rarely accomplishing b oth objectives simultaneously. Here, we address this problem by introducing Rein forcement Learning with Optimistic Ascent-Descent (ReLOAD), a principled CRL met hod with guaranteed last-iterate convergence. We demonstrate its empirical effec tiveness on a wide variety of CRL problems including discrete MDPs and continuou s control. In the process we establish a benchmark of challenging CRL problems.
****

Optimistic Planning by Regularized Dynamic Programming
Antoine Moulin, Gergely Neu
We propose a new method for optimistic planning in infinite-horizon discounted M arkov decision processes based on the idea of adding regularization to the updat es of an otherwise standard approximate value iteration procedure. This techniqu e allows us to avoid contraction and monotonicity arguments typically required b y existing analyses of approximate dynamic programming methods, and in particula r to use approximate transition functions estimated via least-squares procedures in MDPs with linear function approximation. We use our method to recover known guarantees in tabular MDPs and to provide a computationally efficient algorithm for learning near-optimal policies in discounted linear mixture MDPs from a sing le stream of experience, and show it achieves near-optimal statistical guarantee s.
****

Neural signature kernels as infinite-width-depth-limits of controlled ResNets
Nicola Muca Cirone, Maud Lemercier, Cristopher Salvi
Motivated by the paradigm of reservoir computing, we consider randomly initializ ed controlled ResNets defined as Euler-discretizations of neural controlled diff erential equations (Neural CDEs), a unified architecture which enconpasses both RNNs and ResNets. We show that in the infinite-width-depth limit and under prope r scaling, these architectures converge weakly to Gaussian processes indexed on some spaces of continuous paths and with kernels satisfying certain partial diff erential equations (PDEs) varying according to the choice of activation function φ, extending the results of Hayou (2022); Hayou & Yang (2023) to the co ntrolled and homogeneous case. In the special, homogeneous, case where φ is the identity, we show that the equation reduces to a linear PDE and the limi ting kernel agrees with the signature kernel of Salvi et al. (2021a). We name th is new family of limiting kernels neural signature kernels. Finally, we show tha t in the infinite-depth regime, finite-width controlled ResNets converge in dist ribution to Neural CDEs with random vector fields which, depending on whether th e weights are shared across layers, are either time-independent and Gaussian or behave like a matrix-valued Brownian motion.
****

Improving Statistical Fidelity for Neural Image Compression with Implicit Local Likelihood Models
Matthew J. Muckley, Alaaeldin El-Nouby, Karen Ullrich, Herve Jegou, Jakob Verbee

k
Lossy image compression aims to represent images in as few bits as possible whil e maintaining fidelity to the original. Theoretical results indicate that optimi zing distortion metrics such as PSNR or MS-SSIM necessarily leads to a discrepan cy in the statistics of original images from those of reconstructions, in partic ular at low bitrates, often manifested by the blurring of the compressed images. Previous work has leveraged adversarial discriminators to improve statistical f idelity. Yet these binary discriminators adopted from generative modeling tasks may not be ideal for image compression. In this paper, we introduce a non-binary discriminator that is conditioned on quantized local image representations obta ined via VQ-VAE autoencoders. Our evaluations on the CLIC2020, DIV2K and Kodak d atasets show that our discriminator is more effective for jointly optimizing dis tortion (e.g., PSNR) and statistical fidelity (e.g., FID) than the PatchGAN of t he state-of-the-art HiFiC model. On CLIC2020, we obtain the same FID as HiFiC wi th 30-40% fewer bits.
****

PFNs4BO: In-Context Learning for Bayesian Optimization
Samuel Müller, Matthias Feurer, Noah Hollmann, Frank Hutter
In this paper, we use Prior-data Fitted Networks (PFNs) as a flexible surrogate for Bayesian Optimization (BO). PFNs are neural processes that are trained to ap proximate the posterior predictive distribution (PPD) through in-context learnin g on any prior distribution that can be efficiently sampled from. We describe ho w this flexibility can be exploited for surrogate modeling in BO. We use PFNs to mimic a naive Gaussian process (GP), an advanced GP, and a Bayesian Neural Netw ork (BNN). In addition, we show how to incorporate further information into the prior, such as allowing hints about the position of optima (user priors), ignori ng irrelevant dimensions, and performing non-myopic BO by learning the acquisiti on function. The flexibility underlying these extensions opens up vast possibili ties for using PFNs for BO. We demonstrate the usefulness of PFNs for BO in a la rge-scale evaluation on artificial GP samples and three different hyperparameter optimization testbeds: HPO-B, Bayesmark, and PD1. We publish code alongside tra ined models at https://github.com/automl/PFNs4BO.
****

Achieving High Accuracy with PINNs via Energy Natural Gradient Descent
Johannes Müller, Marius Zeinhofer
We propose energy natural gradient descent, a natural gradient method with respe ct to a Hessian-induced Riemannian metric as an optimization algorithm for physi cs-informed neural networks (PINNs) and the deep Ritz method. As a main motivati on we show that the update direction in function space resulting from the energy natural gradient corresponds to the Newton direction modulo an orthogonal proje ction on the model’s tangent space. We demonstrate experimentally that energy na tural gradient descent yields highly accurate solutions with errors several orde rs of magnitude smaller than what is obtained when training PINNs with standard optimizers like gradient descent or Adam, even when those are allowed significan tly more computation time.
****

Uncertain Evidence in Probabilistic Models and Stochastic Simulators
Andreas Munk, Alexander Mead, Frank Wood
We consider the problem of performing Bayesian inference in probabilistic models where observations are accompanied by uncertainty, referred to as "uncertain ev idence.” We explore how to interpret uncertain evidence, and by extension the im portance of proper interpretation as it pertains to inference about latent varia bles. We consider a recently-proposed method "distributional evidence” as well a s revisit two older methods: Jeffrey’s rule and virtual evidence. We devise guid elines on how to account for uncertain evidence and we provide new insights, par ticularly regarding consistency. To showcase the impact of different interpretat ions of the same uncertain evidence, we carry out experiments in which one inter pretation is defined as "correct.” We then compare inference results from each d ifferent interpretation illustrating the importance of careful consideration of uncertain evidence.

****
GibbsDDRM: A Partially Collapsed Gibbs Sampler for Solving Blind Inverse Problems with Denoising Diffusion Restoration
Naoki Murata, Koichi Saito, Chieh-Hsin Lai, Yuhta Takida, Toshimitsu Uesaka, Yuki Mitsufuji, Stefano Ermon
Pre-trained diffusion models have been successfully used as priors in a variety of linear inverse problems, where the goal is to reconstruct a signal from noisy linear measurements. However, existing approaches require knowledge of the linear operator. In this paper, we propose GibbsDDRM, an extension of Denoising Diffusion Restoration Models (DDRM) to a blind setting in which the linear measurement operator is unknown. GibbsDDRM constructs a joint distribution of the data, measurements, and linear operator by using a pre-trained diffusion model for the data prior, and it solves the problem by posterior sampling with an efficient variant of a Gibbs sampler. The proposed method is problem-agnostic, meaning that a pre-trained diffusion model can be applied to various inverse problems without fine-tuning. In experiments, it achieved high performance on both blind image deblurring and vocal dereverberation tasks, despite the use of simple generic priors for the underlying linear operators.
****
DIFF2: Differential Private Optimization via Gradient Differences for Nonconvex Distributed Learning
Tomoya Murata, Taiji Suzuki
Differential private optimization for nonconvex smooth objective is considered. In the previous work, the best known utility bound is $\widetilde O(\sqrt{d}/(n\varepsilon_\mathrm{DP}))$ in terms of the squared full gradient norm, which is achieved by Differential Private Gradient Descent (DP-GD) as an instance, where n is the sample size, d is the problem dimensionality and ε_DP is the differential privacy parameter. To improve the best known utility bound, we propose a new differential private optimization framework called DIFF2 (DIFFerential private optimization via gradient DIFFerences) that constructs a differential private global gradient estimator with possibly quite small variance based on communicated gradient differences rather than gradients themselves. It is shown that DIFF2 with a gradient descent subroutine achieves the utility of $\widetilde O(d^{2/3}/(n\varepsilon_\mathrm{DP})^{4/3})$, which can be significantly better than the previous one in terms of the dependence on the sample size n. To the best of our knowledge, this is the first fundamental result to improve the standard utility $\widetilde O(\sqrt{d}/(n\varepsilon_\mathrm{DP}))$ for nonconvex objectives. Additionally, a more computational and communication efficient subroutine is combined with DIFF2 and its theoretical analysis is also given. Numerical experiments are conducted to validate the superiority of DIFF2 framework.
****
Efficiently predicting high resolution mass spectra with graph neural networks
Michael Murphy, Stefanie Jegelka, Ernest Fraenkel, Tobias Kind, David Healey, Thomas Butler
Identifying a small molecule from its mass spectrum is the primary open problem in computational metabolomics. This is typically cast as information retrieval: an unknown spectrum is matched against spectra predicted computationally from a large database of chemical structures. However, current approaches to spectrum prediction model the output space in ways that force a tradeoff between capturing high resolution mass information and tractable learning. We resolve this tradeoff by casting spectrum prediction as a mapping from an input molecular graph to a probability distribution over chemical formulas. We further discover that a large corpus of mass spectra can be closely approximated using a fixed vocabulary constituting only 2% of all observed formulas. This enables efficient spectrum prediction using an architecture similar to graph classification - GrAFF-MS - achieving significantly lower prediction error and greater retrieval accuracy than previous approaches.
****
Dynamical Linear Bandits

Marco Mussi, Alberto Maria Metelli, Marcello Restelli
In many real-world sequential decision-making problems, an action does not immed iately reflect on the feedback and spreads its effects over a long time frame. F or instance, in online advertising, investing in a platform produces an instanta neous increase of awareness, but the actual reward, i.e., a conversion, might oc cur far in the future. Furthermore, whether a conversion takes place depends on: how fast the awareness grows, its vanishing effects, and the synergy or interfe rence with other advertising platforms. Previous work has investigated the Multi -Armed Bandit framework with the possibility of delayed and aggregated feedback, without a particular structure on how an action propagates in the future, disre garding possible dynamical effects. In this paper, we introduce a novel setting, the Dynamical Linear Bandits (DLB), an extension of the linear bandits characte rized by a hidden state. When an action is performed, the learner observes a noi sy reward whose mean is a linear function of the hidden state and of the action. Then, the hidden state evolves according to linear dynamics, affected by the pe rformed action too. We start by introducing the setting, discussing the notion o f optimal policy, and deriving an expected regret lower bound. Then, we provide an optimistic regret minimization algorithm, Dynamical Linear Upper Confidence B ound (DynLin-UCB), that suffers an expected regret of order $\widetilde{\mathcal {O}} \Big(\frac{d \sqrt{T}}{(1-\overline{\rho})^{3/2}} \Big)$, where $\overline {\rho}$ is a measure of the stability of the system, and d is the dimension of the action vector. Finally, we conduct a numerical validation on a synthetic en vironment and on real-world data to show the effectiveness of DynLin-UCB in comp arison with several baselines.
****
Representation-Driven Reinforcement Learning
Ofir Nabati, Guy Tennenholtz, Shie Mannor
We present a representation-driven framework for reinforcement learning. By repr esenting policies as estimates of their expected values, we leverage techniques from contextual bandits to guide exploration and exploitation. Particularly, emb edding a policy network into a linear feature space allows us to reframe the exp loration-exploitation problem as a representation-exploitation problem, where go od policy representations enable optimal exploration. We demonstrate the effecti veness of this framework through its application to evolutionary and policy grad ient-based approaches, leading to significantly improved performance compared to traditional methods. Our framework provides a new perspective on reinforcement learning, highlighting the importance of policy representation in determining op timal exploration-exploitation strategies.
****
DADAO: Decoupled Accelerated Decentralized Asynchronous Optimization
Adel Nabli, Edouard Oyallon
This work introduces DADAO: the first decentralized, accelerated, asynchronous, primal, first-order algorithm to minimize a sum of L-smooth and μ-strongly convex functions distributed over a given network of size n. Our key insight is based on modeling the local gradient updates and gossip communication procedu res with separate independent Poisson Point Processes. This allows us to decoupl e the computation and communication steps, which can be run in parallel, while m aking the whole approach completely asynchronous. This leads to communication ac celeration compared to synchronous approaches. Our new method employs primal gra dients and does not use a multi-consensus inner loop nor other ad-hoc mechanisms such as Error Feedback, Gradient Tracking, or a Proximal operator. By relating the inverse of the smallest positive eigenvalue of the Laplacian matrix χ_1 and the maximal resistance $\chi_2\leq \chi_1$ of the graph to a sufficient min imal communication rate between the nodes of the network, we show that our algor ithm requires $\mathcal{O}(n\sqrt{\frac{L}{\mu}}\log(\frac{1}{\epsilon}))$ local gradients and only $\mathcal{O}(n\sqrt{\chi_1\chi_2}\sqrt{\frac{L}{\mu}}\log(\f rac{1}{\epsilon}))$ communications to reach a precision ϵ, up to logari thmic terms. Thus, we simultaneously obtain an accelerated rate for both computa tions and communications, leading to an improvement over state-of-the-art works, our simulations further validating the strength of our relatively unconstrained

method.
****
Multi-User Reinforcement Learning with Low Rank Rewards
Dheeraj Mysore Nagaraj, Suhas S Kowshik, Naman Agarwal, Praneeth Netrapalli, Prateek Jain
We consider collaborative multi-user reinforcement learning, where multiple users have the same state-action space and transition probabilities but different rewards. Under the assumption that the reward matrix of the N users has a low-rank structure – a standard and practically successful assumption in the collaborative filtering setting – we design algorithms with significantly lower sample complexity compared to the ones that learn the MDP individually for each user. Our main contribution is an algorithm which explores rewards collaboratively with N user-specific MDPs and can learn rewards efficiently in two key settings: tabular MDPs and linear MDPs. When N is large and the rank is constant, the sample complexity per MDP depends logarithmically over the size of the state-space, which represents an exponential reduction (in the state-space size) when compared to the standard “non-collaborative” algorithms. Our main technical contribution is a method to construct policies which obtain data such that low rank matrix completion is possible (without a generative model). This goes beyond the regular RL framework and is closely related to mean field limits of multi-agent RL.
****
Statistical Foundations of Prior-Data Fitted Networks
Thomas Nagler
Prior-data fitted networks (PFNs) were recently proposed as a new paradigm for machine learning. Instead of training the network to an observed training set, a fixed model is pre-trained offline on small, simulated training sets from a variety of tasks. The pre-trained model is then used to infer class probabilities in-context on fresh training sets with arbitrary size and distribution. Empirically, PFNs achieve state-of-the-art performance on tasks with similar size to the ones used in pre-training. Surprisingly, their accuracy further improves when passed larger data sets during inference. This article establishes a theoretical foundation for PFNs and illuminates the statistical mechanisms governing their behavior. While PFNs are motivated by Bayesian ideas, a purely frequentistic interpretation of PFNs as pre-tuned, but untrained predictors explains their behavior. A predictor’s variance vanishes if its sensitivity to individual training samples does and the bias vanishes only if it is appropriately localized around the test feature. The transformer architecture used in current PFN implementations ensures only the former. These findings shall prove useful for designing architectures with favorable empirical behavior.
****
Do Machine Learning Models Learn Statistical Rules Inferred from Data?
Aaditya Naik, Yinjun Wu, Mayur Naik, Eric Wong
Machine learning models can make critical errors that are easily hidden within vast amounts of data. Such errors often run counter to rules based on human intuition. However, rules based on human knowledge are challenging to scale or to even formalize. We thereby seek to infer statistical rules from the data and quantify the extent to which a model has learned them. We propose a framework SQRL that integrates logic-based methods with statistical inference to derive these rules from a model’s training data without supervision. We further show how to adapt models at test time to reduce rule violations and produce more coherent predictions. SQRL generates up to 300K rules over datasets from vision, tabular, and language settings. We uncover up to 158K violations of those rules by state-of-the-art models for classification, object detection, and data imputation. Test-time adaptation reduces these violations by up to 68.7% with relative performance improvement up to 32%. SQRL is available at https://github.com/DebugML/sqrl.
****
Sample and Predict Your Latent: Modality-free Sequential Disentanglement via Contrastive Estimation
Ilan Naiman, Nimrod Berman, Omri Azencot
Unsupervised disentanglement is a long-standing challenge in representation lear

ning. Recently, self-supervised techniques achieved impressive results in the se
quential setting, where data is time-dependent. However, the latter methods empl
oy modality-based data augmentations and random sampling or solve auxiliary task
s. In this work, we propose to avoid that by generating, sampling, and comparing
 empirical distributions from the underlying variational model. Unlike existing
work, we introduce a self-supervised sequential disentanglement framework based
on contrastive estimation with no external signals, while using common batch siz
es and samples from the latent space itself. In practice, we propose a unified,
efficient, and easy-to-code sampling strategy for semantically similar and dissi
milar views of the data. We evaluate our approach on video, audio, and time seri
es benchmarks. Our method presents state-of-the-art results in comparison to exi
sting techniques. The code is available at https://github.com/azencot-group/SPYL
.
****

Effectively Using Public Data in Privacy Preserving Machine Learning
Milad Nasr, Saeed Mahloujifar, Xinyu Tang, Prateek Mittal, Amir Houmansadr
Differentially private (DP) machine learning techniques are notorious for their
degradation of model utility (e.g., they degrade classification accuracy). A rec
ent line of work has demonstrated that leveraging public data can improve the tr
ade-off between privacy and utility when training models with DP guaranteed. In
this work, we further explore the potential of using public data in DP models, s
howing that utility gains can in fact be significantly higher than what shown in
 prior works. Specifically, we introduce DOPE-SGD, a modified DP-SGD algorithm t
hat leverages public data during its training. DOPE-SGD uses public data in two
complementary ways: (1) it uses advance augmentation techniques that leverages p
ublic data to generate synthetic data that is effectively embedded in multiple s
teps of the training pipeline; (2) it uses a modified gradient clipping mechanis
m (which is a standard technique in DP training) to change the origin of gradien
t vectors using the information inferred from available public and synthetic dat
a, therefore boosting utility. We also introduce a technique to ensemble interme
diate DP models by leveraging the post processing property of differential priva
cy to further improve the accuracy of the predictions. Our experimental results
demonstrate the effectiveness of our approach in improving the state-of-the-art
in DP machine learning across multiple datasets, network architectures, and appl
ication domains. For instance, assuming access to $2,000$ public images, and for
 a privacy budget of $\varepsilon=2,\delta=10^{-5}$, our technique achieves an a
ccuracy of $75.1%$ on CIFAR10, significantly higher than $68.1%$ achieved by the
 state of the art.
****

Counterfactual Identifiability of Bijective Causal Models
Arash Nasr-Esfahany, Mohammad Alizadeh, Devavrat Shah
We study counterfactual identifiability in causal models with bijective generati
on mechanisms (BGM), a class that generalizes several widely-used causal models
in the literature. We establish their counterfactual identifiability for three c
ommon causal structures with unobserved confounding, and propose a practical lea
rning method that casts learning a BGM as structured generative modeling. Learne
d BGMs enable efficient counterfactual estimation and can be obtained using a va
riety of deep conditional generative models. We evaluate our techniques in a vis
ual task and demonstrate its application in a real-world video streaming simulat
ion task.
****

Discovering Object-Centric Generalized Value Functions From Pixels
Somjit Nath, Gopeshh Subbaraj, Khimya Khetarpal, Samira Ebrahimi Kahou
Deep Reinforcement Learning has shown significant progress in extracting useful
representations from high-dimensional inputs albeit using hand-crafted auxiliary
 tasks and pseudo rewards. Automatically learning such representations in an obj
ect-centric manner geared towards control and fast adaptation remains an open re
search problem. In this paper, we introduce a method that tries to discover mean
ingful features from objects, translating them to temporally coherent ‘question’
 functions and leveraging the subsequent learned general value functions for con

trol. We compare our approach with state-of-the-art techniques alongside other a blations and show competitive performance in both stationary and non-stationary settings. Finally, we also investigate the discovered general value functions an d through qualitative analysis show that the learned representations are not onl y interpretable but also, centered around objects that are invariant to changes across tasks facilitating fast adaptation.
****

On Many-Actions Policy Gradient
Michal Nauman, Marek Cygan
We study the variance of stochastic policy gradients (SPGs) with many action sam ples per state. We derive a many-actions optimality condition, which determines when many-actions SPG yields lower variance as compared to a single-action agent with proportionally extended trajectory. We propose Model-Based Many-Actions (M BMA), an approach leveraging dynamics models for many-actions sampling in the co ntext of SPG. MBMA addresses issues associated with existing implementations of many-actions SPG and yields lower bias and comparable variance to SPG estimated from states in model-simulated rollouts. We find that MBMA bias and variance str ucture matches that predicted by theory. As a result, MBMA achieves improved sam ple efficiency and higher returns on a range of continuous action environments a s compared to model-free, many-actions, and model-based on-policy SPG baselines.
****

Equivariant Architectures for Learning in Deep Weight Spaces
Aviv Navon, Aviv Shamsian, Idan Achituve, Ethan Fetaya, Gal Chechik, Haggai Maro n
Designing machine learning architectures for processing neural networks in their raw weight matrix form is a newly introduced research direction. Unfortunately, the unique symmetry structure of deep weight spaces makes this design very chal lenging. If successful, such architectures would be capable of performing a wide range of intriguing tasks, from adapting a pre-trained network to a new domain to editing objects represented as functions (INRs or NeRFs). As a first step tow ards this goal, we present here a novel network architecture for learning in dee p weight spaces. It takes as input a concatenation of weights and biases of a pr e-trained MLP and processes it using a composition of layers that are equivarian t to the natural permutation symmetry of the MLP’s weights: Changing the order o f neurons in intermediate layers of the MLP does not affect the function it repr esents. We provide a full characterization of all affine equivariant and invaria nt layers for these symmetries and show how these layers can be implemented usin g three basic operations: pooling, broadcasting, and fully connected layers appl ied to the input in an appropriate manner. We demonstrate the effectiveness of o ur architecture and its advantages over natural baselines in a variety of learni ng tasks.
****

Scalable Multi-Agent Reinforcement Learning through Intelligent Information Aggr egation
Siddharth Nayak, Kenneth Choi, Wenqi Ding, Sydney Dolan, Karthik Gopalakrishnan, Hamsa Balakrishnan
We consider the problem of multi-agent navigation and collision avoidance when o bservations are limited to the local neighborhood of each agent. We propose Info rMARL, a novel architecture for multi-agent reinforcement learning (MARL) which uses local information intelligently to compute paths for all the agents in a de centralized manner. Specifically, InforMARL aggregates information about the loc al neighborhood of agents for both the actor and the critic using a graph neural network and can be used in conjunction with any standard MARL algorithm. We sho w that (1) in training, InforMARL has better sample efficiency and performance t han baseline approaches, despite using less information, and (2) in testing, it scales well to environments with arbitrary numbers of agents and obstacles. We i llustrate these results using four task environments, including one with predete rmined goals for each agent, and one in which the agents collectively try to cov er all goals.
****

Geometric Autoencoders - What You See is What You Decode
Philipp Nazari, Sebastian Damrich, Fred A Hamprecht
Visualization is a crucial step in exploratory data analysis. One possible appro ach is to train an autoencoder with low-dimensional latent space. Large network depth and width can help unfolding the data. However, such expressive networks c an achieve low reconstruction error even when the latent representation is disto rted. To avoid such misleading visualizations, we propose first a differential g eometric perspective on the decoder, leading to insightful diagnostics for an em bedding’s distortion, and second a new regularizer mitigating such distortion. O ur “Geometric Autoencoder” avoids stretching the embedding spuriously, so that t he visualization captures the data structure more faithfully. It also flags area s where little distortion could not be achieved, thus guarding against misinterp retation.
****

Action Matching: Learning Stochastic Dynamics from Samples
Kirill Neklyudov, Rob Brekelmans, Daniel Severo, Alireza Makhzani
Learning the continuous dynamics of a system from snapshots of its temporal marg inals is a problem which appears throughout natural sciences and machine learnin g, including in quantum systems, single-cell biological data, and generative mod eling. In these settings, we assume access to cross-sectional samples that are u ncorrelated over time, rather than full trajectories of samples. In order to bet ter understand the systems under observation, we would like to learn a model of the underlying process that allows us to propagate samples in time and thereby s imulate entire individual trajectories. In this work, we propose Action Matching , a method for learning a rich family of dynamics using only independent samples from its time evolution. We derive a tractable training objective, which does n ot rely on explicit assumptions about the underlying dynamics and does not requi re back-propagation through differential equations or optimal transport solvers. Inspired by connections with optimal transport, we derive extensions of Action Matching to learn stochastic differential equations and dynamics involving creat ion and destruction of probability mass. Finally, we showcase applications of Ac tion Matching by achieving competitive performance in a diverse set of experimen ts from biology, physics, and generative modeling.
****

Extending Conformal Prediction to Hidden Markov Models with Exact Validity via d e Finetti’s Theorem for Markov Chains
Buddhika Nettasinghe, Samrat Chatterjee, Ramakrishna Tipireddy, Mahantesh M Hala ppanavar
Conformal prediction is a widely used method to quantify the uncertainty of a cl assifier under the assumption of exchangeability (e.g., IID data). We generalize conformal prediction to the Hidden Markov Model (HMM) framework where the assum ption of exchangeability is not valid. The key idea of the proposed method is to partition the non-exchangeable Markovian data from the HMM into exchangeable bl ocks by exploiting the de Finetti’s Theorem for Markov Chains discovered by Diac onis and Freedman (1980). The permutations of the exchangeable blocks are viewed as randomizations of the observed Markovian data from the HMM. The proposed met hod provably retains all desirable theoretical guarantees offered by the classic al conformal prediction framework in both exchangeable and Markovian settings. I n particular, while the lack of exchangeability introduced by Markovian samples constitutes a violation of a crucial assumption for classical conformal predicti on, the proposed method views it as an advantage that can be exploited to improv e the performance further. Detailed numerical and empirical results that complem ent the theoretical conclusions are provided to illustrate the practical feasibi lity of the proposed method.
****

ClimaX: A foundation model for weather and climate
Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K Gupta, Aditya Grover
Recent data-driven approaches based on machine learning aim to directly solve a downstream forecasting or projection task by learning a data-driven functional m apping using deep neural networks. However, these networks are trained using cur

ated and homogeneous climate datasets for specific spatiotemporal tasks, and thu
s lack the generality of currently used computationally intensive physics-inform
ed numerical models for weather and climate modeling. We develop and demonstrate
 ClimaX, a flexible and generalizable deep learning model for weather and climat
e science that can be trained using heterogeneous datasets spanning different va
riables, spatio-temporal coverage, and physical groundings. ClimaX extends the T
ransformer architecture with novel encoding and aggregation blocks that allow ef
fective use of available compute and data while maintaining general utility. Cli
maX is pretrained with a self-supervised learning objective on climate datasets
derived from CMIP6. The pretrained ClimaX can then be fine-tuned to address a br
eadth of climate and weather tasks, including those that involve atmospheric var
iables and spatio-temporal scales unseen during pretraining. Compared to existin
g data-driven baselines, we show that this generality in ClimaX results in super
ior performance on benchmarks for weather forecasting and climate projections, e
ven when pretrained at lower resolutions and compute budgets. Our source code is
 available at https://github.com/microsoft/ClimaX.
****

Provable Reset-free Reinforcement Learning by No-Regret Reduction
Hoai-An Nguyen, Ching-An Cheng
Reinforcement learning (RL) so far has limited real-world applications. One key
challenge is that typical RL algorithms heavily rely on a reset mechanism to sam
ple proper initial states; these reset mechanisms, in practice, are expensive to
 implement due to the need for human intervention or heavily engineered environm
ents. To make learning more practical, we propose a generic no-regret reduction
to systematically design reset-free RL algorithms. Our reduction turns the reset
-free RL problem into a two-player game. We show that achieving sublinear regret
 in this two-player game would imply learning a policy that has both sublinear p
erformance regret and sublinear total number of resets in the original RL proble
m. This means that the agent eventually learns to perform optimally and avoid re
sets. To demonstrate the effectiveness of this reduction, we design an instantia
tion for linear Markov decision processes, which is the first provably correct r
eset-free RL algorithm.
****

Revisiting Over-smoothing and Over-squashing Using Ollivier-Ricci Curvature
Khang Nguyen, Nong Minh Hieu, Vinh Duc Nguyen, Nhat Ho, Stanley Osher, Tan Minh
Nguyen
Graph Neural Networks (GNNs) had been demonstrated to be inherently susceptible
to the problems of over-smoothing and over-squashing. These issues prohibit the
ability of GNNs to model complex graph interactions by limiting their effectiven
ess in taking into account distant information. Our study reveals the key connec
tion between the local graph geometry and the occurrence of both of these issues
, thereby providing a unified framework for studying them at a local scale using
 the Ollivier-Ricci curvature. Specifically, we demonstrate that over-smoothing
is linked to positive graph curvature while over-squashing is linked to negative
 graph curvature. Based on our theory, we propose the Batch Ollivier-Ricci Flow,
 a novel rewiring algorithm capable of simultaneously addressing both over-smoot
hing and over-squashing.
****

Deep Clustering with Incomplete Noisy Pairwise Annotations: A Geometric Regulari
zation Approach
Tri Nguyen, Shahana Ibrahim, Xiao Fu
The recent integration of deep learning and pairwise similarity annotation-based
 constrained clustering—i.e., deep constrained clustering (DCC)—has proven effec
tive for incorporating weak supervision into massive data clustering: Less than
1% of pair similarity annotations can often substantially enhance the clustering
 accuracy. However, beyond empirical successes, there is a lack of understanding
 of DCC. In addition, many DCC paradigms are sensitive to annotation noise, but
performance-guaranteed noisy DCC methods have been largely elusive. This work fi
rst takes a deep look into a recently emerged logistic loss function of DCC, and
 characterizes its theoretical properties. Our result shows that the logistic DC

C loss ensures the identifiability of data membership under reasonable condition
s, which may shed light on its effectiveness in practice. Building upon this und
erstanding, a new loss function based on geometric factor analysis is proposed t
o fend against noisy annotations. It is shown that even under unknown annotation
 confusions, the data membership can still be provably identified under our prop
osed learning criterion. The proposed approach is tested over multiple datasets
to validate our claims.
****
Self-Attention Amortized Distributional Projection Optimization for Sliced Wasse
rstein Point-Cloud Reconstruction
Khai Nguyen, Dang Nguyen, Nhat Ho
Max sliced Wasserstein (Max-SW) distance has been widely known as a solution for
 less discriminative projections of sliced Wasserstein (SW) distance. In applica
tions that have various independent pairs of probability measures, amortized pro
jection optimization is utilized to predict the “max" projecting directions give
n two input measures instead of using projected gradient ascent multiple times.
Despite being efficient, Max-SW and its amortized version cannot guarantee metri
city property due to the sub-optimality of the projected gradient ascent and the
 amortization gap. Therefore, we propose to replace Max-SW with distributional s
liced Wasserstein distance with von Mises-Fisher (vMF) projecting distribution (
v-DSW). Since v-DSW is a metric with any non-degenerate vMF distribution, its am
ortized version can guarantee the metricity when performing amortization. Furthe
rmore, current amortized models are not permutation invariant and symmetric. To
address the issue, we design amortized models based on self-attention architectu
re. In particular, we adopt efficient self-attention architectures to make the c
omputation linear in the number of supports. With the two improvements, we deriv
e self-attention amortized distributional projection optimization and show its a
ppealing performance in point-cloud reconstruction and its downstream applicatio
ns
****
Building Neural Networks on Matrix Manifolds: A Gyrovector Space Approach
Xuan Son Nguyen, Shuo Yang
Matrix manifolds, such as manifolds of Symmetric Positive Definite (SPD) matrice
s and Grassmann manifolds, appear in many applications. Recently, by applying th
e theory of gyrogroups and gyrovector spaces that is a powerful framework for st
udying hyperbolic geometry, some works have attempted to build principled genera
lizations of Euclidean neural networks on matrix manifolds. However, due to the
lack of many concepts in gyrovector spaces for the considered manifolds, e.g., t
he inner product and gyroangles, techniques and mathematical tools provided by t
hese works are still limited compared to those developed for studying hyperbolic
 geometry. In this paper, we generalize some notions in gyrovector spaces for SP
D and Grassmann manifolds, and propose new models and layers for building neural
 networks on these manifolds. We show the effectiveness of our approach in two a
pplications, i.e., human action recognition and knowledge graph completion.
****
Simple Disentanglement of Style and Content in Visual Representations
Lilian Ngweta, Subha Maity, Alex Gittens, Yuekai Sun, Mikhail Yurochkin
Learning visual representations with interpretable features, i.e., disentangled
representations, remains a challenging problem. Existing methods demonstrate som
e success but are hard to apply to large-scale vision datasets like ImageNet. In
 this work, we propose a simple post-processing framework to disentangle content
 and style in learned representations from pre-trained vision models. We model t
he pre-trained features probabilistically as linearly entangled combinations of
the latent content and style factors and develop a simple disentanglement algori
thm based on the probabilistic model. We show that the method provably disentang
les content and style features and verify its efficacy empirically. Our post-pro
cessed features yield significant domain generalization performance improvements
 when the distribution shift occurs due to style changes or style-related spurio
us correlations.
****

MetaDiffuser: Diffusion Model as Conditional Planner for Offline Meta-RL
Fei Ni, Jianye Hao, Yao Mu, Yifu Yuan, Yan Zheng, Bin Wang, Zhixuan Liang
Recently, diffusion model shines as a promising backbone for the sequence modeli ng paradigm in offline reinforcement learning(RL). However, these works mostly l ack the generalization ability across tasks with reward or dynamics change. To t ackle this challenge, in this paper we propose a task-oriented conditioned diffu sion planner for offline meta-RL(MetaDiffuser), which considers the generalizati on problem as conditional trajectory generation task with contextual representat ion. The key is to learn a context conditioned diffusion model which can generat e task-oriented trajectories for planning across diverse tasks. To enhance the d ynamics consistency of the generated trajectories while encouraging trajectories to achieve high returns, we further design a dual-guided module in the sampling process of the diffusion model. The proposed framework enjoys the robustness to the quality of collected warm-start data from the testing task and the flexibil ity to incorporate with different task representation method. The experiment res ults on MuJoCo benchmarks show that MetaDiffuser outperforms other strong offlin e meta-RL baselines, demonstrating the outstanding conditional generation abilit y of diffusion architecture.
****
LEVER: Learning to Verify Language-to-Code Generation with Execution
Ansong Ni, Srini Iyer, Dragomir Radev, Veselin Stoyanov, Wen-Tau Yih, Sida Wang, Xi Victoria Lin
The advent of large language models trained on code (code LLMs) has led to signi ficant progress in language-to-code generation. State-of-the-art approaches in t his area combine LLM decoding with sample pruning and reranking using test cases or heuristics based on the execution results. However, it is challenging to obt ain test cases for many real-world language-to-code applications, and heuristics cannot well capture the semantic features of the execution results, such as dat a type and value range, which often indicates the correctness of the program. In this work, we propose LEVER, a simple approach to improve language-to-code gene ration by learning to verify the generated programs with their execution results . Specifically, we train verifiers to determine whether a program sampled from t he LLMs is correct or not based on the natural language input, the program itsel f and its execution results. The sampled programs are reranked by combining the verification score with the LLM generation probability, and marginalizing over p rograms with the same execution results. On four datasets across the domains of table QA, math QA and basic Python programming, LEVER consistently improves over the base code LLMs (4.6% to 10.9% with code-davinci-002) and achieves new state -of-the-art results on all of them.
****
Continual Vision-Language Representation Learning with Off-Diagonal Information
Zixuan Ni, Longhui Wei, Siliang Tang, Yueting Zhuang, Qi Tian
Large-scale multi-modal contrastive learning frameworks like CLIP typically requ ire a large amount of image-text samples for training. However, these samples ar e always collected continuously in real scenarios. This paper discusses the feas ibility of continual CLIP training using streaming data. Unlike continual learni ng based on self-supervised learning methods for pure images, which is empirical ly robust against catastrophic forgetting, CLIP’s performance degeneration in th e continual setting is significant and non-neglectable. By analyzing the changes in the model’s representation space during continual CLIP training from a spati al geometry perspective, we explore and summarize these spatial variations as Sp atial Disorder (SD), which can be divided into Intra-modal Rotation and Inter-mo dal Deviation. Moreover, we empirically and theoretically demonstrate how SD lea ds to a performance decline for CLIP on cross-modal retrieval tasks. To alleviat e SD, we propose a new continual vision-language representation learning framewo rk Mod-X: Maintain off-diagonal information-matriX. By selectively aligning the off-diagonal information distribution of contrastive matrices, the Mod-X improve s the capability of the multi-modal model by maintaining the multi-modal represe ntation space alignment on the old data domain during continuously fitting the n ew training data domain. Experiments on commonly used datasets with different sc

ales and scopes have demonstrated the effectiveness of our method.
****
Attributing Image Generative Models using Latent Fingerprints
Guangyu Nie, Changhoon Kim, Yezhou Yang, Yi Ren
Generative models have enabled the creation of contents that are indistinguishable from those taken from nature. Open-source development of such models raised concerns about the risks of their misuse for malicious purposes. One potential risk mitigation strategy is to attribute generative models via fingerprinting. Current fingerprinting methods exhibit a significant tradeoff between robust attribution accuracy and generation quality while lacking design principles to improve this tradeoff. This paper investigates the use of latent semantic dimensions as fingerprints, from where we can analyze the effects of design variables, including the choice of fingerprinting dimensions, strength, and capacity, on the accuracy-quality tradeoff. Compared with previous SOTA, our method requires minimum computation and is more applicable to large-scale models. We use StyleGAN2 and the latent diffusion model to demonstrate the efficacy of our method.
****
A Framework for Adapting Offline Algorithms to Solve Combinatorial Multi-Armed Bandit Problems with Bandit Feedback
Guanyu Nie, Yididiya Y. Nadew, Yanhui Zhu, Vaneet Aggarwal, Christopher John Quinn
We investigate the problem of stochastic, combinatorial multi-armed bandits where the learner only has access to bandit feedback and the reward function can be non-linear. We provide a general framework for adapting discrete offline approximation algorithms into sublinear α-regret methods that only require bandit feedback, achieving $\mathcal{O}\left(T^\frac{2}{3}\log(T)^\frac{1}{3}\right)$ expected cumulative α-regret dependence on the horizon T. The framework only requires the offline algorithms to be robust to small errors in function evaluation. The adaptation procedure does not even require explicit knowledge of the offline approximation algorithm — the offline algorithm can be used as black box subroutine. To demonstrate the utility of the proposed framework, the proposed framework is applied to multiple problems in submodular maximization, adapting approximation algorithms for cardinality and for knapsack constraints. The new CMAB algorithms for knapsack constraints outperform a full-bandit method developed for the adversarial setting in experiments with real-world data.
****
SinFusion: Training Diffusion Models on a Single Image or Video
Yaniv Nikankin, Niv Haim, Michal Irani
Diffusion models exhibited tremendous progress in image and video generation, exceeding GANs in quality and diversity. However, they are usually trained on very large datasets and are not naturally adapted to manipulate a given input image or video. In this paper we show how this can be resolved by training a diffusion model on a single input image or video. Our image/video-specific diffusion model (SinFusion) learns the appearance and dynamics of the single image or video, while utilizing the conditioning capabilities of diffusion models. It can solve a wide array of image/video-specific manipulation tasks. In particular, our model can learn from few frames the motion and dynamics of a single input video. It can then generate diverse new video samples of the same dynamic scene, extrapolate short videos into long ones (both forward and backward in time) and perform video upsampling. Most of these tasks are not realizable by current video-specific generation methods.
****
SparseProp: Efficient Sparse Backpropagation for Faster Training of Neural Networks at the Edge
Mahdi Nikdan, Tommaso Pegolotti, Eugenia Iofinova, Eldar Kurtic, Dan Alistarh
We provide an efficient implementation of the backpropagation algorithm, specialized to the case where the weights of the neural network being trained are sparse. Our algorithm is general, as it applies to arbitrary (unstructured) sparsity and common layer types (e.g., convolutional or linear). We provide a fast vectorized implementation on commodity CPUs, and show that it can yield speedups in en

d-to-end runtime experiments, both in transfer learning using already-sparsified networks, and in training sparse networks from scratch. Thus, our results provide the first support for sparse training on commodity hardware.
****

Anti-Exploration by Random Network Distillation
Alexander Nikulin, Vladislav Kurenkov, Denis Tarasov, Sergey Kolesnikov
Despite the success of Random Network Distillation (RND) in various domains, it was shown as not discriminative enough to be used as an uncertainty estimator for penalizing out-of-distribution actions in offline reinforcement learning. In this paper, we revisit these results and show that, with a naive choice of conditioning for the RND prior, it becomes infeasible for the actor to effectively minimize the anti-exploration bonus and discriminativity is not an issue. We show that this limitation can be avoided with conditioning based on Feature-wise Linear Modulation (FiLM), resulting in a simple and efficient ensemble-free algorithm based on Soft Actor-Critic. We evaluate it on the D4RL benchmark, showing that it is capable of achieving performance comparable to ensemble-based methods and outperforming ensemble-free approaches by a wide margin.
****

Input Perturbation Reduces Exposure Bias in Diffusion Models
Mang Ning, Enver Sangineto, Angelo Porrello, Simone Calderara, Rita Cucchiara
Denoising Diffusion Probabilistic Models have shown an impressive generation quality although their long sampling chain leads to high computational costs. In this paper, we observe that a long sampling chain also leads to an error accumulation phenomenon, which is similar to the exposure bias problem in autoregressive text generation. Specifically, we note that there is a discrepancy between training and testing, since the former is conditioned on the ground truth samples, while the latter is conditioned on the previously generated results. To alleviate this problem, we propose a very simple but effective training regularization, consisting in perturbing the ground truth samples to simulate the inference time prediction errors. We empirically show that, without affecting the recall and precision, the proposed input perturbation leads to a significant improvement in the sample quality while reducing both the training and the inference times. For instance, on CelebA 64x64, we achieve a new state-of-the-art FID score of 1.27, while saving 37.5% of the training time. The code is available at https://github.com/forever208/DDPM-IP
****

Primal and Dual Analysis of Entropic Fictitious Play for Finite-sum Problems
Atsushi Nitanda, Kazusato Oko, Denny Wu, Nobuhito Takenouchi, Taiji Suzuki
The entropic fictitious play (EFP) is a recently proposed algorithm that minimizes the sum of a convex functional and entropy in the space of measures — such an objective naturally arises in the optimization of a two-layer neural network in the mean-field regime. In this work, we provide a concise primal-dual analysis of EFP in the setting where the learning problem exhibits a finite-sum structure. We establish quantitative global convergence guarantees for both the continuous-time and discrete-time dynamics based on properties of a proximal Gibbs measure introduced in Nitanda et al. (2022). Furthermore, our primal-dual framework entails a memory-efficient particle-based implementation of the EFP update, and also suggests a connection to gradient boosting methods. We illustrate the efficiency of our novel implementation in experiments including neural network optimization and image synthesis.
****

The Statistical Scope of Multicalibration
Georgy Noarov, Aaron Roth
We make a connection between multicalibration and property elicitation and show that (under mild technical conditions) it is possible to produce a multicalibrated predictor for a continuous scalar property Γ if and only if Γ is elicitable. On the negative side, we show that for non-elicitable continuous properties there exist simple data distributions on which even the true distributional predictor is not calibrated. On the positive side, for elicitable Γ, we give simple canonical algorithms for the batch and the online adversarial s

etting, that learn a Γ-multicalibrated predictor. This generalizes past work on multicalibrated means and quantiles, and in fact strengthens existing online quantile multicalibration results. To further counter-weigh our negative result, we show that if a property Γ^1 is not elicitable by itself, but is elicitable conditionally on another elicitable property Γ^0, then there is a canonical algorithm that jointly multicalibrates Γ^1 and Γ^0; this generalizes past work on mean-moment multicalibration. Finally, as applications of our theory, we provide novel algorithmic and impossibility results for fair (multicalibrated) risk assessment.
****

Do Embodied Agents Dream of Pixelated Sheep: Embodied Decision Making using Language Guided World Modelling
Kolby Nottingham, Prithviraj Ammanabrolu, Alane Suhr, Yejin Choi, Hannaneh Hajishirzi, Sameer Singh, Roy Fox
Reinforcement learning (RL) agents typically learn tabula rasa, without prior knowledge of the world. However, if initialized with knowledge of high-level subgoals and transitions between subgoals, RL agents could utilize this Abstract World Model (AWM) for planning and exploration. We propose using few-shot large language models (LLMs) to hypothesize an AWM, that will be verified through world experience, to improve sample efficiency of RL agents. Our DECKARD agent applies LLM-guided exploration to item crafting in Minecraft in two phases: (1) the Dream phase where the agent uses an LLM to decompose a task into a sequence of subgoals, the hypothesized AWM; and (2) the Wake phase where the agent learns a modular policy for each subgoal and verifies or corrects the hypothesized AWM. Our method of hypothesizing an AWM with LLMs and then verifying the AWM based on agent experience not only increases sample efficiency over contemporary methods by an order of magnitude but is also robust to and corrects errors in the LLM, successfully blending noisy internet-scale information from LLMs with knowledge grounded in environment dynamics.
****

Gradient-Free Structured Pruning with Unlabeled Data
Azade Nova, Hanjun Dai, Dale Schuurmans
Large Language Models (LLMs) have achieved great success in solving difficult tasks across many domains, but such success comes with a high computation cost, and inference latency. As developers and third parties customize these models, the need to provide efficient inference has increased. Many efforts have attempted to reduce inference cost through model compression techniques such as pruning and distillation. However, these techniques either require labeled data, or are time-consuming as they require the compressed model to be retrained to regain accuracy. In this paper, we propose a gradient-free structured pruning framework that uses only unlabeled data. An evaluation on the GLUE and SQuAD benchmarks using BERT$_{BASE}$ and DistilBERT illustrates the effectiveness of the proposed approach. By only using the weights of the pre-trained model and unlabeled data, in a matter of a few minutes on a single GPU, up to 40% of the original FLOP count can be reduced with less than a $4%$ accuracy loss across all tasks considered.
****

CHiLS: Zero-Shot Image Classification with Hierarchical Label Sets
Zachary Novack, Julian Mcauley, Zachary Chase Lipton, Saurabh Garg
Open vocabulary models (e.g. CLIP) have shown strong performance on zero-shot classification through their ability generate embeddings for each class based on their (natural language) names. Prior work has focused on improving the accuracy of these models through prompt engineering or by incorporating a small amount of labeled downstream data (via finetuning). However, there has been little focus on improving the richness of the class names themselves, which can pose issues when class labels are coarsely-defined and are uninformative. We propose Classification with Hierarchical Label Sets (or CHiLS), an alternative strategy for zero-shot classification specifically designed for datasets with implicit semantic hierarchies. CHiLS proceeds in three steps: (i) for each class, produce a set of subclasses, using either existing label hierarchies or by querying GPT-3; (ii) perform the standard zero-shot CLIP procedure as though these subclasses were the

labels of interest; (iii) map the predicted subclass back to its parent to produce the final prediction. Across numerous datasets with underlying hierarchical structure, CHiLS leads to improved accuracy in situations both with and without ground-truth hierarchical information. CHiLS is simple to implement within existing zero-shot pipelines and requires no additional training cost. Code is available at: https://github.com/acmi-lab/CHILS.
****

Few-bit Backward: Quantized Gradients of Activation Functions for Memory Footprint Reduction
Georgii Sergeevich Novikov, Daniel Bershatsky, Julia Gusak, Alex Shonenkov, Denis Valerievich Dimitrov, Ivan Oseledets
Memory footprint is one of the main limiting factors for large neural network training. In backpropagation, one needs to store the input to each operation in the computational graph. Every modern neural network model has quite a few pointwise nonlinearities in its architecture, and such operations induce additional memory costs that, as we show, can be significantly reduced by quantization of the gradients. We propose a systematic approach to compute optimal quantization of the retained gradients of the pointwise nonlinear functions with only a few bits per each element. We show that such approximation can be achieved by computing an optimal piecewise-constant approximation of the derivative of the activation function, which can be done by dynamic programming. The drop-in replacements are implemented for all popular nonlinearities and can be used in any existing pipeline. We confirm the memory reduction and the same convergence on several open benchmarks.
****

Efficient Exploration via Epistemic-Risk-Seeking Policy Optimization
Brendan O’Donoghue
Exploration remains a key challenge in deep reinforcement learning (RL). Optimism in the face of uncertainty is a well-known heuristic with theoretical guarantees in the tabular setting, but how best to translate the principle to deep reinforcement learning, which involves online stochastic gradients and deep network function approximators, is not fully understood. In this paper we propose a new, differentiable optimistic objective that when optimized yields a policy that provably explores efficiently, with guarantees even under function approximation. Our new objective is a zero-sum two-player game derived from endowing the agent with an epistemic-risk-seeking utility function, which converts uncertainty into value and encourages the agent to explore uncertain states. We show that the solution to this game minimizes an upper bound on the regret, with the ’players’ each attempting to minimize one component of a particular regret decomposition. We derive a new model-free algorithm which we call ’epistemic-risk-seeking actor-critic’ (ERSAC), which is simply an application of simultaneous stochastic gradient ascent-descent to the game. Finally, we discuss a recipe for incorporating off-policy data and show that combining the risk-seeking objective with replay data yields a double benefit in terms of statistical efficiency. We conclude with some results showing good performance of a deep RL agent using the technique on the challenging ’DeepSea’ environment, showing significant performance improvements even over other efficient exploration techniques, as well as improved performance on the Atari benchmark.
****

Provable Benefit of Mixup for Finding Optimal Decision Boundaries
Junsoo Oh, Chulhee Yun
We investigate how pair-wise data augmentation techniques like Mixup affect the sample complexity of finding optimal decision boundaries in a binary linear classification problem. For a family of data distributions with a separability constant κ, we analyze how well the optimal classifier in terms of training loss aligns with the optimal one in test accuracy (i.e., Bayes optimal classifier). For vanilla training without augmentation, we uncover an interesting phenomenon named the curse of separability. As we increase κ to make the data distribution more separable, the sample complexity of vanilla training increases exponentially in κ; perhaps surprisingly, the task of finding optimal decisi

on boundaries becomes harder for more separable distributions. For Mixup trainin g, we show that Mixup mitigates this problem by significantly reducing the sampl e complexity. To this end, we develop new concentration results applicable to $n ^2$ pair-wise augmented data points constructed from n independent data, by ca refully dealing with dependencies between overlapping pairs. Lastly, we study ot her masking-based Mixup-style techniques and show that they can distort the trai ning loss and make its minimizer converge to a suboptimal classifier in terms of test accuracy.

****

Shedding a PAC-Bayesian Light on Adaptive Sliced-Wasserstein Distances

Ruben Ohana, Kimia Nadjahi, Alain Rakotomamonjy, Liva Ralaivola

The Sliced-Wasserstein distance (SW) is a computationally efficient and theoreti cally grounded alternative to the Wasserstein distance. Yet, the literature on i ts statistical properties – or, more accurately, its generalization properties – with respect to the distribution of slices, beyond the uniform measure, is scar ce. To bring new contributions to this line of research, we leverage the PAC-Bay esian theory and a central observation that SW may be interpreted as an average risk, the quantity PAC-Bayesian bounds have been designed to characterize. We pr ovide three types of results: i) PAC-Bayesian generalization bounds that hold on what we refer as adaptive Sliced-Wasserstein distances, i.e. SW defined with re spect to arbitrary distributions of slices (among which data-dependent distribut ions), ii) a principled procedure to learn the distribution of slices that yield s maximally discriminative SW, by optimizing our theoretical bounds, and iii) em pirical illustrations of our theoretical findings.

****

Reasons for the Superiority of Stochastic Estimators over Deterministic Ones: Ro bustness, Consistency and Perceptual Quality

Guy Ohayon, Theo Joseph Adrai, Michael Elad, Tomer Michaeli

Stochastic restoration algorithms allow to explore the space of solutions that c orrespond to the degraded input. In this paper we reveal additional fundamental advantages of stochastic methods over deterministic ones, which further motivate their use. First, we prove that any restoration algorithm that attains perfect perceptual quality and whose outputs are consistent with the input must be a pos terior sampler, and is thus required to be stochastic. Second, we illustrate tha t while deterministic restoration algorithms may attain high perceptual quality, this can be achieved only by filling up the space of all possible source images using an extremely sensitive mapping, which makes them highly vulnerable to adv ersarial attacks. Indeed, we show that enforcing deterministic models to be robu st to such attacks profoundly hinders their perceptual quality, while robustifyi ng stochastic models hardly influences their perceptual quality, and improves th eir output variability. These findings provide a motivation to foster progress i n stochastic restoration methods, paving the way to better recovery algorithms.

****

On the Within-Group Fairness of Screening Classifiers

Nastaran Okati, Stratis Tsirtsis, Manuel Gomez Rodriguez

Screening classifiers are increasingly used to identify qualified candidates in a variety of selection processes. In this context, it has been recently shown th at if a classifier is calibrated, one can identify the smallest set of candidate s which contains, in expectation, a desired number of qualified candidates using a threshold decision rule. This lends support to focusing on calibration as the only requirement for screening classifiers. In this paper, we argue that screen ing policies that use calibrated classifiers may suffer from an understudied typ e of within-group unfairness—they may unfairly treat qualified members within de mographic groups of interest. Further, we argue that this type of unfairness can be avoided if classifiers satisfy within-group monotonicity, a natural monotoni city property within each group. Then, we introduce an efficient post-processing algorithm based on dynamic programming to minimally modify a given calibrated c lassifier so that its probability estimates satisfy within-group monotonicity. W e validate our algorithm using US Census survey data and show that within-group monotonicity can often be achieved at a small cost in terms of prediction granul

arity and shortlist size.
****
Diffusion Models are Minimax Optimal Distribution Estimators
Kazusato Oko, Shunta Akiyama, Taiji Suzuki
While efficient distribution learning is no doubt behind the groundbreaking success of diffusion modeling, its theoretical guarantees are quite limited. In this paper, we provide the first rigorous analysis on approximation and generalization abilities of diffusion modeling for well-known function spaces. The highlight of this paper is that when the true density function belongs to the Besov space and the empirical score matching loss is properly minimized, the generated data distribution achieves the nearly minimax optimal estimation rates in the total variation distance and in the Wasserstein distance of order one. Furthermore, we extend our theory to demonstrate how diffusion models adapt to low-dimensional data distributions. We expect these results advance theoretical understandings of diffusion modeling and its ability to generate verisimilar outputs.
****
How Many Perturbations Break This Model? Evaluating Robustness Beyond Adversarial Accuracy
Raphael Olivier, Bhiksha Raj
Robustness to adversarial attacks is typically evaluated with adversarial accuracy. While essential, this metric does not capture all aspects of robustness and in particular leaves out the question of how many perturbations can be found for each point. In this work, we introduce an alternative approach, adversarial sparsity, which quantifies how difficult it is to find a successful perturbation given both an input point and a constraint on the direction of the perturbation. We show that sparsity provides valuable insight into neural networks in multiple ways: for instance, it illustrates important differences between current state-of-the-art robust models them that accuracy analysis does not, and suggests approaches for improving their robustness. When applying broken defenses effective against weak attacks but not strong ones, sparsity can discriminate between the totally ineffective and the partially effective defenses. Finally, with sparsity we can measure increases in robustness that do not affect accuracy: we show for example that data augmentation can by itself increase adversarial robustness, without using adversarial training.
****
B-Learner: Quasi-Oracle Bounds on Heterogeneous Causal Effects Under Hidden Confounding
Miruna Oprescu, Jacob Dorn, Marah Ghoummaid, Andrew Jesson, Nathan Kallus, Uri Shalit
Estimating heterogeneous treatment effects from observational data is a crucial task across many fields, helping policy and decision-makers take better actions. There has been recent progress on robust and efficient methods for estimating the conditional average treatment effect (CATE) function, but these methods often do not take into account the risk of hidden confounding, which could arbitrarily and unknowingly bias any causal estimate based on observational data. We propose a meta-learner called the B-Learner, which can efficiently learn sharp bounds on the CATE function under limits on the level of hidden confounding. We derive the B-Learner by adapting recent results for sharp and valid bounds of the average treatment effect (Dorn et al., 2021) into the framework given by Kallus & Oprescu (2023) for robust and model-agnostic learning of conditional distributional treatment effects. The B-Learner can use any function estimator such as random forests and deep neural networks, and we prove its estimates are valid, sharp, efficient, and have a quasi-oracle property with respect to the constituent estimators under more general conditions than existing methods. Semi-synthetic experimental comparisons validate the theoretical findings, and we use real-world data demonstrate how the method might be used in practice.
****
Measuring the Impact of Programming Language Distribution
Gabriel Orlanski, Kefan Xiao, Xavier Garcia, Jeffrey Hui, Joshua Howland, Jonathan Malmaud, Jacob Austin, Rishabh Singh, Michele Catasta

Current benchmarks for evaluating neural code models focus on only a small subset of programming languages, excluding many popular languages such as Go or Rust. To ameliorate this issue, we present the BabelCode framework for execution-based evaluation of any benchmark in any language. BabelCode enables new investigations into the qualitative performance of models' memory, runtime, and individual test case results. Additionally, we present a new code translation dataset called Translating Python Programming Puzzles (TP3) from the Python Programming Puzzles (Schuster et al., 2021) benchmark that involves translating expert-level python functions to any language. With both BabelCode and the TP3 benchmark, we investigate if balancing the distributions of 14 languages in a training dataset improves a large language model's performance on low-resource languages. Training a model on a balanced corpus results in, on average, 12.34% higher $pass@k$ across all tasks and languages compared to the baseline. We find that this strategy achieves 66.48% better $pass@k$ on low-resource languages at the cost of only a 12.94% decrease to high-resource languages. In our three translation tasks, this strategy yields, on average, 30.77% better low-resource $pass@k$ while having 19.58% worse high-resource $pass@k$.

****

When does Privileged information Explain Away Label Noise?

Guillermo Ortiz-Jimenez, Mark Collier, Anant Nawalgaria, Alexander Nicholas D'Amour, Jesse Berent, Rodolphe Jenatton, Efi Kokiopoulou

Leveraging privileged information (PI), or features available during training but not at test time, has recently been shown to be an effective method for addressing label noise. However, the reasons for its effectiveness are not well understood. In this study, we investigate the role played by different properties of the PI in explaining away label noise. Through experiments on multiple datasets with real PI (CIFAR-N/H) and a new large-scale benchmark ImageNet-PI, we find that PI is most helpful when it allows networks to easily distinguish clean from noisy data, while enabling a learning shortcut to memorize the noisy examples. Interestingly, when PI becomes too predictive of the target label, PI methods often perform worse than their no-PI baselines. Based on these findings, we propose several enhancements to the state-of-the-art PI methods and demonstrate the potential of PI as a means of tackling label noise. Finally, we show how we can easily combine the resulting PI approaches with existing no-PI techniques designed to deal with label noise.

****

Resurrecting Recurrent Neural Networks for Long Sequences

Antonio Orvieto, Samuel L Smith, Albert Gu, Anushan Fernando, Caglar Gulcehre, Razvan Pascanu, Soham De

Recurrent Neural Networks (RNNs) offer fast inference on long sequences but are hard to optimize and slow to train. Deep state-space models (SSMs) have recently been shown to perform remarkably well on long sequence modeling tasks, and have the added benefits of fast parallelizable training and RNN-like fast inference. However, while SSMs are superficially similar to RNNs, there are important differences that make it unclear where their performance boost over RNNs comes from. We show that careful design of deep RNNs using standard signal propagation arguments can recover the impressive performance of deep SSMs on long-range reasoning tasks, while matching their training speed. To achieve this, we analyze and ablate a series of changes to standard RNNs including linearizing and diagonalizing the recurrence, using better parameterizations and initializations, and ensuring careful normalization of the forward pass. Our results provide new insights on the origins of the impressive performance of deep SSMs, and introduce an RNN block called the Linear Recurrent Unit (or LRU) that matches both their performance on the Long Range Arena benchmark and their computational efficiency.

****

Improving Adversarial Robustness Through the Contrastive-Guided Diffusion Process

Yidong Ouyang, Liyan Xie, Guang Cheng

Synthetic data generation has become an emerging tool to help improve the adversarial robustness in classification tasks, since robust learning requires a signi

ficantly larger amount of training samples compared with standard classification . Among various deep generative models, the diffusion model has been shown to pr oduce high-quality synthetic images and has achieved good performance in improvi ng the adversarial robustness. However, diffusion-type methods are generally slo wer in data generation as compared with other generative models. Although differ ent acceleration techniques have been proposed recently, it is also of great imp ortance to study how to improve the sample efficiency of synthetic data for the downstream task. In this paper, we first analyze the optimality condition of syn thetic distribution for achieving improved robust accuracy. We show that enhanci ng the distinguishability among the generated data is critical for improving adv ersarial robustness. Thus, we propose the Contrastive-Guided Diffusion Process (Contrastive-DP), which incorporates the contrastive loss to guide the diffusion model in data generation. We validate our theoretical results using simulations and demonstrate the good performance of Contrastive-DP on image datasets.

****

On the Role of Attention in Prompt-tuning

Samet Oymak, Ankit Singh Rawat, Mahdi Soltanolkotabi, Christos Thrampoulidis

Prompt-tuning is an emerging strategy to adapt large language models (LLM) to do wnstream tasks by learning a (soft-)prompt parameter from data. Despite its succ ess in LLMs, there is limited theoretical understanding of the power of prompt-t uning and the role of the attention mechanism in prompting. In this work, we exp lore prompt-tuning for one-layer attention architectures and study contextual mi xture-models where each input token belongs to a context-relevant or -irrelevant set. We isolate the role of prompt-tuning through a self-contained prompt-atten tion model. Our contributions are as follows: (1) We show that softmax-prompt-at tention is provably more expressive than softmax-self-attention and linear-promp t-attention under our contextual data model. (2) We analyze the initial trajecto ry of gradient descent and show that it learns the prompt and prediction head wi th near-optimal sample complexity and demonstrate how the prompt can provably at tend to sparse context-relevant tokens. (3) Assuming a known prompt but an unkno wn prediction head, we characterize the exact finite sample performance of promp t-attention which reveals the fundamental performance limits and the precise ben efit of the context information. We also provide experiments that verify our the oretical insights on real datasets and demonstrate how prompt-tuning enables the model to attend to context-relevant information.

****

Revisiting the Linear-Programming Framework for Offline RL with General Function Approximation

Asuman E. Ozdaglar, Sarath Pattathil, Jiawei Zhang, Kaiqing Zhang

Offline reinforcement learning (RL) aims to find an optimal policy for sequentia l decision-making using a pre-collected dataset, without further interaction wit h the environment. Recent theoretical progress has focused on developing sample- efficient offline RL algorithms with various relaxed assumptions on data coverag e and function approximators, especially to handle the case with excessively lar ge state-action spaces. Among them, the framework based on the linear-programmin g (LP) reformulation of Markov decision processes has shown promise: it enables sample-efficient offline RL with function approximation, under only partial data coverage and realizability assumptions on the function classes, with favorable computational tractability. In this work, we revisit the LP framework for offlin e RL, and provide a new reformulation that advances the existing results in seve ral aspects, relaxing certain assumptions and achieving optimal statistical rate s in terms of sample size. Our key enabler is to introduce proper constraints in the reformulation, instead of using any regularization as in the literature, al so with careful choices of the function classes and initial state distributions. We hope our insights bring into light the use of LP formulations and the induce d primal-dual minimax optimization, in offline RL.

****

Extrapolative Controlled Sequence Generation via Iterative Refinement

Vishakh Padmakumar, Richard Yuanzhe Pang, He He, Ankur P Parikh

We study the problem of extrapolative controlled generation, i.e., generating se

quences with attribute values beyond the range seen in training. This task is of significant importance in automated design, especially drug discovery, where the goal is to design novel proteins that are better (e.g., more stable) than existing sequences. Thus, by definition the target sequences and their attribute values are out of the training distribution, posing challenges to existing methods that aim to directly generate the target sequence. Instead, in this work, we propose Iterative Controlled Extrapolation (ICE) which iteratively makes local edits to a sequence to enable extrapolation. We train the model on synthetically generated sequence pairs that demonstrate small improvement in the attribute value. Results on one natural language task (sentiment analysis) and two protein engineering tasks (ACE2 stability and AAV fitness) show that ICE outperforms state-of-the-art approaches despite its simplicity.

****

Locally Regularized Neural Differential Equations: Some Black Boxes were meant to remain closed!

Avik Pal, Alan Edelman, Christopher Vincent Rackauckas

Neural Differential Equations have become an important modeling framework due to their ability to adapt to new problems automatically. Training a neural differential equation is effectively a search over a space of plausible dynamical systems. Controlling the computational cost for these models is difficult since it relies on the number of steps the adaptive solver takes. Most prior works have used higher-order methods to reduce prediction timings while greatly increasing training time or reducing both training and prediction timings by relying on specific training algorithms, which are harder to use as a drop-in replacement. In this manuscript, we use internal cost heuristics of adaptive differential equation solvers at stochastic time-points to guide the training towards learning a dynamical system that is easier to integrate. We “close the blackbox” and allow the use of our method with any sensitivity method. We perform experimental studies to compare our method to global regularization to show that we attain similar performance numbers without compromising on the flexibility of implementation. We develop two sampling strategies to trade-off between performance and training time. Our method reduces the number of function evaluations to 0.556x - 0.733x and accelerates predictions by 1.3x - 2x.

****

Controlled Differential Equations on Long Sequences via Non-standard Wavelets

Sourav Pal, Zhanpeng Zeng, Sathya N. Ravi, Vikas Singh

Neural Controlled Differential equations (NCDE) are a powerful mechanism to model the dynamics in temporal sequences, e.g., applications involving physiological measures, where apart from the initial condition, the dynamics also depend on subsequent measures or even a different "control" sequence. But NCDEs do not scale well to longer sequences. Existing strategies adapt rough path theory, and instead model the dynamics over summaries known as log signatures. While rigorous and elegant, invertibility of these summaries is difficult, and limits the scope of problems where these ideas can offer strong benefits (reconstruction, generative modeling). For tasks where it is sensible to assume that the (long) sequences in the training data are a fixed length of temporal measurements – this assumption holds in most experiments tackled in the literature – we describe an efficient simplification. First, we recast the regression/classification task as an integral transform. We then show how restricting the class of operators (permissible in the integral transform), allows the use of a known algorithm that leverages non-standard Wavelets to decompose the operator. Thereby, our task (learning the operator) radically simplifies. A neural variant of this idea yields consistent improvements across a wide gamut of use cases tackled in existing works. We also describe a novel application on modeling tasks involving coupled differential equations.

****

Do the Rewards Justify the Means? Measuring Trade-Offs Between Rewards and Ethical Behavior in the Machiavelli Benchmark

Alexander Pan, Jun Shern Chan, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside, Hanlin Zhang, Scott Emmons, Dan Hendrycks

Artificial agents have traditionally been trained to maximize reward, which may incentivize power-seeking and deception, analogous to how next-token prediction in language models (LMs) may incentivize toxicity. So do agents naturally learn to be Machiavellian? And how do we measure these behaviors in general-purpose models such as GPT-4? Towards answering these questions, we introduce Machiavelli, a benchmark of 134 Choose-Your-Own-Adventure games containing over half a million rich, diverse scenarios that center on social decision-making. Scenario labeling is automated with LMs, which are more performant than human annotators. We mathematize dozens of harmful behaviors and use our annotations to evaluate agents' tendencies to be power-seeking, cause disutility, and commit ethical violations. We observe some tension between maximizing reward and behaving ethically. To improve this trade-off, we investigate LM-based methods to steer agents towards less harmful behaviors. Our results show that agents can both act competently and morally, so concrete progress can currently be made in machine ethics–designing agents that are Pareto improvements in both safety and capabilities.
****
Beyond Homophily: Reconstructing Structure for Graph-agnostic Clustering
Erlin Pan, Zhao Kang
Graph neural networks (GNNs) based methods have achieved impressive performance on node clustering task. However, they are designed on the homophilic assumption of graph and clustering on heterophilic graph is overlooked. Due to the lack of labels, it is impossible to first identify a graph as homophilic or heterophilic before a suitable GNN model can be found. Hence, clustering on real-world graph with various levels of homophily poses a new challenge to the graph research community. To fill this gap, we propose a novel graph clustering method, which contains three key components: graph reconstruction, a mixed filter, and dual graph clustering network. To be graph-agnostic, we empirically construct two graphs which are high homophily and heterophily from each data. The mixed filter based on the new graphs extracts both low-frequency and high-frequency information. To reduce the adverse coupling between node attribute and topological structure, we separately map them into two subspaces in dual graph clustering network. Extensive experiments on 11 benchmark graphs demonstrate our promising performance. In particular, our method dominates others on heterophilic graphs.
****
Better Training of GFlowNets with Local Credit and Incomplete Trajectories
Ling Pan, Nikolay Malkin, Dinghuai Zhang, Yoshua Bengio
Generative Flow Networks or GFlowNets are related to Monte-Carlo Markov chain methods (as they sample from a distribution specified by an energy function), reinforcement learning (as they learn a policy to sample composed objects through a sequence of steps), generative models (as they learn to represent and sample from a distribution) and amortized variational methods (as they can be used to learn to approximate and sample from an otherwise intractable posterior, given a prior and a likelihood). They are trained to generate an object x through a sequence of steps with probability proportional to some reward function $R(x)$ (or $\exp(-\mathcal{E}(x))$ with $\mathcal{E}(x)$ denoting the energy function), given at the end of the generative trajectory. Like for other RL settings where the reward is only given at the end, the efficiency of training and credit assignment may suffer when those trajectories are longer. With previous GFlowNet work, no learning was possible from incomplete trajectories (lacking a terminal state and the computation of the associated reward). In this paper, we consider the case where the energy function can be applied not just to terminal states but also to intermediate states. This is for example achieved when the energy function is additive, with terms available along the trajectory. We show how to reparameterize the GFlowNet state flow function to take advantage of the partial reward already accrued at each state. This enables a training objective that can be applied to update parameters even with incomplete trajectories. Even when complete trajectories are available, being able to obtain more localized credit and gradients is found to speed up training convergence, as demonstrated across many simulations.
****

A Hybrid Quantum-Classical Approach based on the Hadamard Transform for the Convolutional Layer
Hongyi Pan, Xin Zhu, Salih Furkan Atici, Ahmet Cetin
In this paper, we propose a novel Hadamard Transform (HT)-based neural network layer for hybrid quantum-classical computing. It implements the regular convolutional layers in the Hadamard transform domain. The idea is based on the HT convolution theorem which states that the dyadic convolution between two vectors is equivalent to the element-wise multiplication of their HT representation. Computing the HT is simply the application of a Hadamard gate to each qubit individually, so the HT computations of our proposed layer can be implemented on a quantum computer. Compared to the regular Conv2D layer, the proposed HT-perceptron layer is computationally more efficient. Compared to a CNN with the same number of trainable parameters and 99.26% test accuracy, our HT network reaches 99.31% test accuracy with 57.1% MACs reduced in the MNIST dataset; and in our ImageNet-1K experiments, our HT-based ResNet-50 exceeds the accuracy of the baseline ResNet-50 by 0.59% center-crop top-1 accuracy using 11.5% fewer parameters with 12.6% fewer MACs.
****
Semi Bandit dynamics in Congestion Games: Convergence to Nash Equilibrium and No-Regret Guarantees.
Ioannis Panageas, Stratis Skoulakis, Luca Viano, Xiao Wang, Volkan Cevher
In this work, we propose introduce a variant of online stochastic gradient descent and prove it converges to Nash equilibria and simultaneously it has sublinear regret for the class of congestion games in the semi-bandit feedback setting. Our proposed method admits convergence rates depending only polynomially on the number of players and the number of facilities, but not on the size of the action set, which can be exponentially large in terms of the number of facilities. Moreover, the running time of our method has polynomial-time dependence on the implicit description of the game. Our analysis exploits techniques from convex geometry, in particular Caratheodory’s theorem and recent advances in non-convex stochastic optimization. This work improves upon and answers an open question from (Cui et al 2022).
****
Flash: Concept Drift Adaptation in Federated Learning
Kunjal Panchal, Sunav Choudhary, Subrata Mitra, Koyel Mukherjee, Somdeb Sarkhel, Saayan Mitra, Hui Guan
In Federated Learning (FL), adaptive optimization is an effective approach to addressing the statistical heterogeneity issue but cannot adapt quickly to concept drifts. In this work, we propose a novel adaptive optimizer called Flash that simultaneously addresses both statistical heterogeneity and the concept drift issues. The fundamental insight is that a concept drift can be detected based on the magnitude of parameter updates that are required to fit the global model to each participating client’s local data distribution. Flash uses a two-pronged approach that synergizes client-side early-stopping training to facilitate detection of concept drifts and the server-side drift-aware adaptive optimization to effectively adjust effective learning rate. We theoretically prove that Flash matches the convergence rate of state-of-the-art adaptive optimizers and further empirically evaluate the efficacy of Flash on a variety of FL benchmarks using different concept drift settings.
****
Learn to Accumulate Evidence from All Training Samples: Theory and Practice
Deep Shankar Pandey, Qi Yu
Evidential deep learning, built upon belief theory and subjective logic, offers a principled and computationally efficient way to turn a deterministic neural network uncertainty-aware. The resultant evidential models can quantify fine-grained uncertainty using the learned evidence. To ensure theoretically sound evidential models, the evidence needs to be non-negative, which requires special activation functions for model training and inference. This constraint often leads to inferior predictive performance compared to standard softmax models, making it challenging to extend them to many large-scale datasets. To unveil the real cause

of this undesired behavior, we theoretically investigate evidential models and identify a fundamental limitation that explains the inferior performance: existing evidential activation functions create zero evidence regions, which prevent the model to learn from training samples falling into such regions. A deeper analysis of evidential activation functions based on our theoretical underpinning inspires the design of a novel regularizer that effectively alleviates this fundamental limitation. Extensive experiments over many challenging real-world datasets and settings confirm our theoretical findings and demonstrate the effectiveness of our proposed approach.
****
Secure Federated Correlation Test and Entropy Estimation
Qi Pang, Lun Wang, Shuai Wang, Wenting Zheng, Dawn Song
We propose the first federated correlation test framework compatible with secure aggregation, namely FED-χ^2. In our protocol, the statistical computations are recast as frequency moment estimation problems, where the clients collaboratively generate a shared projection matrix and then use stable projection to encode the local information in a compact vector. As such encodings can be linearly aggregated, secure aggregation can be applied to conceal the individual updates. We formally establish the security guarantee of FED-χ^2 by proving that only the minimum necessary information (i.e., the correlation statistics) is revealed to the server. We show that our protocol can be naturally extended to estimate other statistics that can be recast as frequency moment estimations. By accommodating Shannon’e Entropy in FED-χ^2, we further propose the first secure federated entropy estimation protocol, FED-H. The evaluation results demonstrate that FED-χ^2 and FED-H achieve good performance with small client-side computation overhead in several real-world case studies.
****
Task-Specific Skill Localization in Fine-tuned Language Models
Abhishek Panigrahi, Nikunj Saunshi, Haoyu Zhao, Sanjeev Arora
Pre-trained language models can be fine-tuned to solve diverse NLP tasks, including in few-shot settings. Thus fine-tuning allows the model to quickly pick up task-specific "skills," but there has been limited study of where these newly-learnt skills reside inside the massive model. This paper introduces the term skill localization for this problem and proposes a solution. Given the downstream task and a model fine-tuned on that task, a simple optimization is used to identify a very small subset of parameters ($\sim$0.01% of model parameters) responsible for ($>$95%) of the model’s performance, in the sense that grafting the fine-tuned values for just this tiny subset onto the pre-trained model gives performance almost as well as the fine-tuned model. While reminiscent of recent works on parameter-efficient fine-tuning, the novel aspects here are that: (i) No further retraining is needed on the subset (unlike, say, with lottery tickets). (ii) Notable improvements are seen over vanilla fine-tuning with respect to calibration of predictions in-distribution (40-90% error reduction) as well as quality of predictions out-of-distribution (OOD). In models trained on multiple tasks, a stronger notion of skill localization is observed, where the sparse regions corresponding to different tasks are almost disjoint, and their overlap (when it happens) is a proxy for task similarity. Experiments suggest that localization via grafting can assist certain forms continual learning.
****
Kernel Sufficient Dimension Reduction and Variable Selection for Compositional Data via Amalgamation
Junyoung Park, Jeongyoun Ahn, Cheolwoo Park
Compositional data with a large number of components and an abundance of zeros are frequently observed in many fields recently. Analyzing such sparse high-dimensional compositional data naturally calls for dimension reduction or, more preferably, variable selection. Most existing approaches lack interpretability or cannot handle zeros properly, as they rely on a log-ratio transformation. We approach this problem with sufficient dimension reduction (SDR), one of the most studied dimension reduction frameworks in statistics. Characterized by the conditional independence of the data to the response on the found subspace, the SDR framew

ork has been effective for both linear and nonlinear dimension reduction problems. This work proposes a compositional SDR that can handle zeros naturally while incorporating the nonlinear nature and spurious negative correlations among components rigorously. A critical consideration of sub-composition versus amalgamation for compositional variable selection is discussed. The proposed compositional SDR is shown to be statistically consistent in constructing a sub-simplex consisting of true signal variables. Simulation and real microbiome data are used to demonstrate the performance of the proposed SDR compared to existing state-of-art approaches.

****

Learning Affinity with Hyperbolic Representation for Spatial Propagation

Jin-Hwi Park, Jaesung Choe, Inhwan Bae, Hae-Gon Jeon

Recent approaches to representation learning have successfully demonstrated the benefits in hyperbolic space, driven by an excellent ability to make hierarchical relationships. In this work, we demonstrate that the properties of hyperbolic geometry serve as a valuable alternative to learning hierarchical affinity for spatial propagation tasks. We propose a Hyperbolic Affinity learning Module (HAM) to learn spatial affinity by considering geodesic distance on the hyperbolic space. By simply incorporating our HAM into conventional spatial propagation tasks, we validate its effectiveness, capturing the pixel hierarchy of affinity maps in hyperbolic space. The proposed methodology can lead to performance improvements in explicit propagation processes such as depth completion and semantic segmentation.

****

TRAK: Attributing Model Behavior at Scale

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, Aleksander Madry

The goal of data attribution is to trace model predictions back to training data. Despite a long line of work towards this goal, existing approaches to data attribution tend to force users to choose between computational tractability and efficacy. That is, computationally tractable methods can struggle with accurately attributing model predictions in non-convex settings (e.g., in the context of deep neural networks), while methods that are effective in such regimes require training thousands of models, which makes them impractical for large models or datasets. In this work, we introduce TRAK (Tracing with the Randomly-projected After Kernel), a data attribution method that is both effective and computationally tractable for large-scale, differentiable models. In particular, by leveraging only a handful of trained models, TRAK can match the performance of attribution methods that require training thousands of models. We demonstrate the utility of TRAK across various modalities and scales: image classifiers trained on ImageNet, vision-language models (CLIP), and language models (BERT and mT5). We provide code for using TRAK (and reproducing our work) at https://github.com/MadryLab/trak .

****

Test-Time Style Shifting: Handling Arbitrary Styles in Domain Generalization

Jungwuk Park, Dong-Jun Han, Soyeong Kim, Jaekyun Moon

In domain generalization (DG), the target domain is unknown when the model is being trained, and the trained model should successfully work on an arbitrary (and possibly unseen) target domain during inference. This is a difficult problem, and despite active studies in recent years, it remains a great challenge. In this paper, we take a simple yet effective approach to tackle this issue. We propose test-time style shifting, which shifts the style of the test sample (that has a large style gap with the source domains) to the nearest source domain that the model is already familiar with, before making the prediction. This strategy enables the model to handle any target domains with arbitrary style statistics, without additional model update at test-time. Additionally, we propose style balancing, which provides a great platform for maximizing the advantage of test-time style shifting by handling the DG-specific imbalance issues. The proposed ideas are easy to implement and successfully work in conjunction with various other DG schemes. Experimental results on different datasets show the effectiveness of our

methods.
****
Towards Understanding Ensemble Distillation in Federated Learning
Sejun Park, Kihun Hong, Ganguk Hwang
Federated Learning (FL) is a collaborative machine learning paradigm for data privacy preservation. Recently, a knowledge distillation (KD) based information sharing approach in FL, which conducts ensemble distillation on an unlabeled public dataset, has been proposed. However, despite its experimental success and usefulness, the theoretical analysis of the KD based approach has not been satisfactorily conducted. In this work, we build a theoretical foundation of the ensemble distillation framework in federated learning from the perspective of kernel ridge regression (KRR). In this end, we propose a KD based FL algorithm for KRR models which is related with some existing KD based FL algorithms, and analyze our algorithm theoretically. We show that our algorithm makes local prediction models as much powerful as the centralized KRR model (which is a KRR model trained by all of local datasets) in terms of the convergence rate of the generalization error if the unlabeled public dataset is sufficiently large. We also provide experimental results to verify our theoretical results on ensemble distillation in federated learning.
****
Learning Controllable Degradation for Real-World Super-Resolution via Constrained Flows
Seobin Park, Dongjin Kim, Sungyong Baik, Tae Hyun Kim
Recent deep-learning-based super-resolution (SR) methods have been successful in recovering high-resolution (HR) images from their low-resolution (LR) counterparts, albeit on the synthetic and simple degradation setting: bicubic downscaling. On the other hand, super-resolution on real-world images demands the capability to handle complex downscaling mechanism which produces different artifacts (e.g., noise, blur, color distortion) upon downscaling factors. To account for complex downscaling mechanism in real-world LR images, there have been a few efforts in constructing datasets consisting of LR images with real-world downsampling degradation. However, making such datasets entails a tremendous amount of time and effort, thereby resorting to very few number of downscaling factors (e.g., $\times$2, $\times$3, $\times$4). To remedy the issue, we propose to generate realistic SR datasets for unseen degradation levels by exploring the latent space of real LR images and thereby producing more diverse yet realistic LR images with complex real-world artifacts. Our quantitative and qualitative experiments demonstrate the accuracy of the generated LR images, and we show that the various conventional SR networks trained with our newly generated SR datasets can produce much better HR images.
****
Differentially Private Sharpness-Aware Training
Jinseong Park, Hoki Kim, Yujin Choi, Jaewook Lee
Training deep learning models with differential privacy (DP) results in a degradation of performance. The training dynamics of models with DP show a significant difference from standard training, whereas understanding the geometric properties of private learning remains largely unexplored. In this paper, we investigate sharpness, a key factor in achieving better generalization, in private learning. We show that flat minima can help reduce the negative effects of per-example gradient clipping and the addition of Gaussian noise. We then verify the effectiveness of Sharpness-Aware Minimization (SAM) for seeking flat minima in private learning. However, we also discover that SAM is detrimental to the privacy budget and computational time due to its two-step optimization. Thus, we propose a new sharpness-aware training method that mitigates the privacy-optimization trade-off. Our experimental results demonstrate that the proposed method improves the performance of deep learning models with DP from both scratch and fine-tuning. Code is available at https://github.com/jinseongP/DPSAT.
****
Controllability-Aware Unsupervised Skill Discovery
Seohong Park, Kimin Lee, Youngwoon Lee, Pieter Abbeel

One of the key capabilities of intelligent agents is the ability to discover useful skills without external supervision. However, the current unsupervised skill discovery methods are often limited to acquiring simple, easy-to-learn skills due to the lack of incentives to discover more complex, challenging behaviors. We introduce a novel unsupervised skill discovery method, Controllability-aware Skill Discovery (CSD), which actively seeks complex, hard-to-control skills without supervision. The key component of CSD is a controllability-aware distance function, which assigns larger values to state transitions that are harder to achieve with the current skills. Combined with distance-maximizing skill discovery, CSD progressively learns more challenging skills over the course of training as our jointly trained distance function reduces rewards for easy-to-achieve skills. Our experimental results in six robotic manipulation and locomotion environments demonstrate that CSD can discover diverse complex skills including object manipulation and locomotion skills with no supervision, significantly outperforming prior unsupervised skill discovery methods. Videos and code are available at https://seohong.me/projects/csd/
****

Predictable MDP Abstraction for Unsupervised Model-Based RL
Seohong Park, Sergey Levine
A key component of model-based reinforcement learning (RL) is a dynamics model that predicts the outcomes of actions. Errors in this predictive model can degrade the performance of model-based controllers, and complex Markov decision processes (MDPs) can present exceptionally difficult prediction problems. To mitigate this issue, we propose predictable MDP abstraction (PMA): instead of training a predictive model on the original MDP, we train a model on a transformed MDP with a learned action space that only permits predictable, easy-to-model actions, while covering the original state-action space as much as possible. As a result, model learning becomes easier and more accurate, which allows robust, stable model-based planning or model-based RL. This transformation is learned in an unsupervised manner, before any task is specified by the user. Downstream tasks can then be solved with model-based control in a zero-shot fashion, without additional environment interactions. We theoretically analyze PMA and empirically demonstrate that PMA leads to significant improvements over prior unsupervised model-based RL approaches in a range of benchmark environments. Our code and videos are available at https://seohong.me/projects/pma/
****

Neural Stochastic Differential Games for Time-series Analysis
Sungwoo Park, Byoungwoo Park, Moontae Lee, Changhee Lee
Modeling spatiotemporal dynamics with neural differential equations has become a major line of research that opens new ways to handle various real-world scenarios (e.g., missing observations, irregular times, etc.). Despite such progress, most existing methods still face challenges in providing a general framework for analyzing time series. To tackle this, we adopt stochastic differential games to suggest a new philosophy of utilizing interacting collective intelligence in time series analysis. For the implementation, we develop the novel gradient descent-based algorithm called deep neural fictitious play to approximate the Nash equilibrium. We theoretically analyze the convergence result of the proposed algorithm and discuss the advantage of cooperative games in handling noninformative observation. Throughout the experiments on various datasets, we demonstrate the superiority of our framework over all the tested benchmarks in modeling time-series prediction by capitalizing on the advantages of applying cooperative games. An ablation study shows that neural agents of the proposed framework learn intrinsic temporal relevance to make accurate time-series predictions.
****

Accelerated Infeasibility Detection of Constrained Optimization and Fixed-Point Iterations
Jisun Park, Ernest K. Ryu
As first-order optimization methods become the method of choice for solving large-scale optimization problems, optimization solvers based on first-order algorithms are being built. Such general-purpose solvers must robustly detect infeasibl

e or misspecified problem instances, but the computational complexity of first-o rder methods for doing so has yet to be formally studied. In this work, we chara cterize the optimal accelerated rate of infeasibility detection. We show that th e standard fixed-point iteration achieves a $\mathcal{O}(1/k^2)$ and $\mathcal{O }(1/k)$ rates, respectively, on the normalized iterates and the fixed-point resi dual converging to the infimal displacement vector, while the accelerated fixed-point iteration achieves $\mathcal{O}(1/k^2)$ and $\tilde{\mathcal{O}}(1/k^2)$ r ates. We then provide a matching complexity lower bound to establish that $\Thet a(1/k^2)$ is indeed the optimal accelerated rate.
****

Model-based Reinforcement Learning with Scalable Composite Policy Gradient Estim ators

Paavo Parmas, Takuma Seno, Yuma Aoki

In model-based reinforcement learning (MBRL), policy gradients can be estimated either by derivative-free RL methods, such as likelihood ratio gradients (LR), o r by backpropagating through a differentiable model via reparameterization gradi ents (RP). Instead of using one or the other, the Total Propagation (TP) algorit hm in prior work showed that a combination of LR and RP estimators averaged usin g inverse variance weighting (IVW) can achieve orders of magnitude improvement o ver either method. However, IVW-based composite estimators have not yet been app lied in modern RL tasks, as it is unclear if they can be implemented scalably. W e propose a scalable method, Total Propagation X (TPX) that improves over TP by changing the node used for IVW, and employing coordinate wise weighting. We demo nstrate the scalability of TPX by applying it to the state of the art visual MBR L algorithm Dreamer. The experiments showed that Dreamer fails with long simulat ion horizons, while our TPX works reliably for only a fraction of additional com putation. One key advantage of TPX is its ease of implementation, which will ena ble experimenting with IVW on many tasks beyond MBRL.
****

PAC Generalization via Invariant Representations

Advait U Parulekar, Karthikeyan Shanmugam, Sanjay Shakkottai

Invariant representations are transformations of the covariates such that the be st model on top of the representation is invariant across training environments. In the context of linear Structural Equation Models (SEMs), invariant represent ations might allow us to learn models with out-of-distribution guarantees, i.e., models that are robust to interventions in the SEM. To address the invariant re presentation problem in a finite sample setting, we consider the notion of $\eps ilon$-approximate invariance. We study the following question: If a representati on is approximately invariant with respect to a given number of training interve ntions, will it continue to be approximately invariant on a larger collection of unseen intervened SEMs? Inspired by PAC learning, we obtain finite-sample out-o f-distribution generalization guarantees for approximate invariance that holds p robabilistically over a family of linear SEMs without faithfulness assumptions.
****

Stochastic Gradient Descent-Induced Drift of Representation in a Two-Layer Neura l Network

Farhad Pashakhanloo, Alexei Koulakov

Representational drift refers to over-time changes in neural activation accompan ied by a stable task performance. Despite being observed in the brain and in art ificial networks, the mechanisms of drift and its implications are not fully und erstood. Motivated by recent experimental findings of stimulus-dependent drift i n the piriform cortex, we use theory and simulations to study this phenomenon in a two-layer linear feedforward network. Specifically, in a continual online lea rning scenario, we study the drift induced by the noise inherent in the Stochast ic Gradient Descent (SGD). By decomposing the learning dynamics into the normal and tangent spaces of the minimum-loss manifold, we show the former corresponds to a finite variance fluctuation, while the latter could be considered as an eff ective diffusion process on the manifold. We analytically compute the fluctuatio n and the diffusion coefficients for the stimuli representations in the hidden l ayer as functions of network parameters and input distribution. Further, consist

ent with experiments, we show that the drift rate is slower for a more frequentl y presented stimulus. Overall, our analysis yields a theoretical framework for b etter understanding of the drift phenomenon in biological and artificial neural networks.

****

Reducing SO(3) Convolutions to SO(2) for Efficient Equivariant GNNs

Saro Passaro, C. Lawrence Zitnick

Graph neural networks that model 3D data, such as point clouds or atoms, are typ ically desired to be $SO(3)$ equivariant, i.e., equivariant to 3D rotations. Unf ortunately equivariant convolutions, which are a fundamental operation for equiv ariant networks, increase significantly in computational complexity as higher-or der tensors are used. In this paper, we address this issue by reducing the $SO(3)$ convolutions or tensor products to mathematically equivalent convolutions in $SO(2)$. This is accomplished by aligning the node embeddings' primary axis wit h the edge vectors, which sparsifies the tensor product and reduces the computat ional complexity from $O(L^6)$ to $O(L^3)$, where L is the degree of the repre sentation. We demonstrate the potential implications of this improvement by prop osing the Equivariant Spherical Channel Network (eSCN), a graph neural network u tilizing our novel approach to equivariant convolutions, which achieves state-of -the-art results on the large-scale OC-20 and OC-22 datasets.

****

Federated Online and Bandit Convex Optimization

Kumar Kshitij Patel, Lingxiao Wang, Aadirupa Saha, Nathan Srebro

We study the problems of distributed online and bandit convex optimization again st an adaptive adversary. We aim to minimize the average regret on M machines working in parallel over T rounds with R intermittent communications. Assumi ng the underlying cost functions are convex and can be generated adaptively, our results show that collaboration is not beneficial when the machines have access to the first-order gradient information at the queried points. This is in contr ast to the case for stochastic functions, where each machine samples the cost fu nctions from a fixed distribution. Furthermore, we delve into the more challengi ng setting of federated online optimization with bandit (zeroth-order) feedback, where the machines can only access values of the cost functions at the queried points. The key finding here is identifying the high-dimensional regime where co llaboration is beneficial and may even lead to a linear speedup in the number of machines. We further illustrate our findings through federated adversarial line ar bandits by developing novel distributed single and two-point feedback algorit hms. Our work is the first attempt towards a systematic understanding of federat ed online optimization with limited feedback, and it attains tight regret bounds in the intermittent communication setting for both first and zeroth-order feedb ack. Our results thus bridge the gap between stochastic and adaptive settings in federated online optimization.

****

Brauer's Group Equivariant Neural Networks

Edward Pearce-Crump

We provide a full characterisation of all of the possible group equivariant neur al networks whose layers are some tensor power of $\mathbb{R}^{n}$ for three sym metry groups that are missing from the machine learning literature: $O(n)$, the orthogonal group; $SO(n)$, the special orthogonal group; and $Sp(n)$, the symple ctic group. In particular, we find a spanning set of matrices for the learnable, linear, equivariant layer functions between such tensor power spaces in the sta ndard basis of $\mathbb{R}^{n}$ when the group is $O(n)$ or $SO(n)$, and in the symplectic basis of $\mathbb{R}^{n}$ when the group is $Sp(n)$.

****

How Jellyfish Characterise Alternating Group Equivariant Neural Networks

Edward Pearce-Crump

We provide a full characterisation of all of the possible alternating group ($A_ n$) equivariant neural networks whose layers are some tensor power of $\mathbb{R }^{n}$. In particular, we find a basis of matrices for the learnable, linear, $A _n$–equivariant layer functions between such tensor power spaces in the standard

basis of $\mathbb{R}^{n}$. We also describe how our approach generalises to the construction of neural networks that are equivariant to local symmetries.
****

Can Large Language Models Reason about Program Invariants?
Kexin Pei, David Bieber, Kensen Shi, Charles Sutton, Pengcheng Yin
Identifying invariants is an important program analysis task with applications towards program understanding, bug finding, vulnerability analysis, and formal verification. Existing tools for identifying program invariants rely on dynamic analysis, requiring traces collected from multiple executions in order to produce reliable invariants. We study the application of large language models to invariant prediction, finding that models trained on source code and fine-tuned for invariant generation can perform invariant prediction as static rather than dynamic analysis. Using a scratchpad approach where invariants are predicted sequentially through a program gives the best performance, finding invariants statically of quality comparable to those obtained by a dynamic analysis tool with access to five program traces.
****

Dynamics-inspired Neuromorphic Visual Representation Learning
Zhengqi Pei, Shuhui Wang
This paper investigates the dynamics-inspired neuromorphic architecture for visual representation learning following Hamilton’s principle. Our method converts weight-based neural structure to its dynamics-based form that consists of finite sub-models, whose mutual relations measured by computing path integrals amongst their dynamical states are equivalent to the typical neural weights. Based on the entropy reduction process derived from the Euler-Lagrange equations, the feedback signals interpreted as stress forces amongst sub-models push them to move. We first train a dynamics-based neural model from scratch and observe that this model outperforms traditional neural models on MNIST. We then convert several pre-trained neural structures into dynamics-based forms, followed by fine-tuning via entropy reduction to obtain the stabilized dynamical states. We observe consistent improvements in these transformed models over their weight-based counterparts on ImageNet and WebVision in terms of computational complexity, parameter size, testing accuracy, and robustness. Besides, we show the correlation between model performance and structural entropy, providing deeper insight into weight-free neuromorphic learning.
****

Feature Directions Matter: Long-Tailed Learning via Rotated Balanced Representation
Gao Peifeng, Qianqian Xu, Peisong Wen, Zhiyong Yang, Huiyang Shao, Qingming Huang
Long-tailed learning is one of the most challenging problems in visual recognition. There are some studies aiming to solve long-tailed classification from the perspective of feature learning. Recent work proposes to learn the balanced representation by fixing the linear classifier as Equiangular Tight Frame (ETF), since they argue what matters in classification is the structure of the feature, instead of their directions. Holding a different view, in this paper, we show that features with fixed directions may be harmful to the generalization of models, even if it is completely symmetric. To avoid this issue, we propose Representation-Balanced Learning Framework (RBL), which introduces orthogonal matrices to learn directions while maintaining the geometric structure of ETF. Theoretically, our contributions are two-fold: 1). we point out that the feature learning of RBL is insensitive toward training set label distribution, it always learns a balanced representation space. 2). we provide a generalization analysis of proposed RBL through training stability. To analyze the stability of the parameter with orthogonal constraint, we propose a novel training stability analysis paradigm, Two-Parameter Model Stability. Practically, our method is extremely simple in implementation but shows great superiority on several benchmark datasets.
****

Fair Neighbor Embedding
Jaakko Peltonen, Wen Xu, Timo Nummenmaa, Jyrki Nummenmaa

We consider fairness in dimensionality reduction. Nonlinear dimensionality reduc
tion yields low dimensional representations that let users visualize and explore
 high-dimensional data. However, traditional dimensionality reduction may yield
biased visualizations overemphasizing relationships of societal phenomena to sen
sitive attributes or protected groups. We introduce a framework of fair neighbor
 embedding, the Fair Neighbor Retrieval Visualizer, which formulates fair nonlin
ear dimensionality reduction as an information retrieval task whose performance
and fairness are quantified by information retrieval criteria. The method optimi
zes low-dimensional embeddings that preserve high-dimensional data neighborhoods
 without yielding biased association of such neighborhoods to protected groups.
In experiments the method yields fair visualizations outperforming previous meth
ods.
****

The Ideal Continual Learner: An Agent That Never Forgets
Liangzu Peng, Paris Giampouras, Rene Vidal
The goal of continual learning is to find a model that solves multiple learning
tasks which are presented sequentially to the learner. A key challenge in this s
etting is that the learner may "forget" how to solve a previous task when learni
ng a new task, a phenomenon known as catastrophic forgetting. To address this ch
allenge, many practical methods have been proposed, including memory-based, regu
larization-based and expansion-based methods. However, a rigorous theoretical un
derstanding of these methods remains elusive. This paper aims to bridge this gap
 between theory and practice by proposing a new continual learning framework cal
led "Ideal Continual Learner" (ICL), which is guaranteed to avoid catastrophic f
orgetting by construction. We show that ICL unifies multiple well-established co
ntinual learning methods and gives new theoretical insights into the strengths a
nd weaknesses of these methods. We also derive generalization bounds for ICL whi
ch allow us to theoretically quantify "how rehearsal affects generalization". Fi
nally, we connect ICL to several classic subjects and research topics of modern
interest, which allows us to make historical remarks and inspire future directio
ns.
****

MolDiff: Addressing the Atom-Bond Inconsistency Problem in 3D Molecule Diffusion
 Generation
Xingang Peng, Jiaqi Guan, Qiang Liu, Jianzhu Ma
Deep generative models have recently achieved superior performance in 3D molecul
e generation. Most of them first generate atoms and then add chemical bonds base
d on the generated atoms in a post-processing manner. However, there might be no
 corresponding bond solution for the temporally generated atoms as their locatio
ns are generated without considering potential bonds. We define this problem as
the atom-bond inconsistency problem and claim it is the main reason for current
approaches to generating unrealistic 3D molecules. To overcome this problem, we
propose a new diffusion model called MolDiff which can generate atoms and bonds
simultaneously while still maintaining their consistency by explicitly modeling
the dependence between their relationships. We evaluated the generation ability
of our proposed model and the quality of the generated molecules using criteria
related to both geometry and chemical properties. The empirical studies showed t
hat our model outperforms previous approaches, achieving a three-fold improvemen
t in success rate and generating molecules with significantly better quality.
****

Diagnosis, Feedback, Adaptation: A Human-in-the-Loop Framework for Test-Time Pol
icy Adaptation
Andi Peng, Aviv Netanyahu, Mark K Ho, Tianmin Shu, Andreea Bobu, Julie Shah, Pul
kit Agrawal
Policies often fail at test-time due to distribution shifts—changes in the state
 and reward that occur when an end user deploys the policy in environments diffe
rent from those seen in training. Data augmentation can help models be more robu
st to such shifts by varying specific concepts in the state, e.g. object color,
that are task-irrelevant and should not impact desired actions. However, designe
rs training the agent don’t often know which concepts are irrelevant a priori. W

e propose a human-in-the-loop framework to leverage feedback from the end user t
o quickly identify and augment task-irrelevant visual state concepts. Our framew
ork generates counterfactual demonstrations that allow users to quickly isolate
shifted state concepts and identify if they should not impact the desired task,
and can therefore be augmented using existing actions. We present experiments va
lidating our full pipeline on discrete and continuous control tasks with real hu
man users. Our method better enables users to (1) understand agent failure, (2)
improve sample efficiency of demonstrations required for finetuning, and (3) ada
pt the agent to their desired reward.
****

Learning Hidden Markov Models When the Locations of Missing Observations are Unk
nown

Binyamin Perets, Mark Kozdoba, Shie Mannor

The Hidden Markov Model (HMM) is one of the most widely used statistical models
for sequential data analysis. One of the key reasons for this versatility is the
 ability of HMM to deal with missing data. However, standard HMM learning algori
thms rely crucially on the assumption that the positions of the missing observat
ions within the observation sequence are known. In the natural sciences, where t
his assumption is often violated, special variants of HMM, commonly known as Sil
ent-state HMMs (SHMMs), are used. Despite their widespread use, these algorithms
 strongly rely on specific structural assumptions of the underlying chain, such
as acyclicity, thus limiting the applicability of these methods. Moreover, even
in the acyclic case, it has been shown that these methods can lead to poor recon
struction. In this paper we consider the general problem of learning an HMM from
 data with unknown missing observation locations. We provide reconstruction algo
rithms that do not require any assumptions about the structure of the underlying
 chain, and can also be used with limited prior knowledge, unlike SHMM. We evalu
ate and compare the algorithms in a variety of scenarios, measuring their recons
truction precision, and robustness under model miss-specification. Notably, we s
how that under proper specifications one can reconstruct the process dynamics as
 well as if the missing observations positions were known.
****

Estimating the Contamination Factor’s Distribution in Unsupervised Anomaly Detec
tion

Lorenzo Perini, Paul-Christian Bürkner, Arto Klami

Anomaly detection methods identify examples that do not follow the expected beha
viour, typically in an unsupervised fashion, by assigning real-valued anomaly sc
ores to the examples based on various heuristics. These scores need to be transf
ormed into actual predictions by thresholding so that the proportion of examples
 marked as anomalies equals the expected proportion of anomalies, called contami
nation factor. Unfortunately, there are no good methods for estimating the conta
mination factor itself. We address this need from a Bayesian perspective, introd
ucing a method for estimating the posterior distribution of the contamination fa
ctor for a given unlabeled dataset. We leverage several anomaly detectors to cap
ture the basic notion of anomalousness and estimate the contamination using a sp
ecific mixture formulation. Empirically on 22 datasets, we show that the estimat
ed distribution is well-calibrated and that setting the threshold using the post
erior mean improves the detectors’ performance over several alternative methods.
****

Are Gaussian Data All You Need? The Extents and Limits of Universality in High-D
imensional Generalized Linear Estimation

Luca Pesce, Florent Krzakala, Bruno Loureiro, Ludovic Stephan

In this manuscript we consider the problem of generalized linear estimation on G
aussian mixture data with labels given by a single-index model. Our first result
 is a sharp asymptotic expression for the test and training errors in the high-d
imensional regime. Motivated by the recent stream of results on the Gaussian uni
versality of the test and training errors in generalized linear estimation, we a
sk ourselves the question: "when is a single Gaussian enough to characterize the
 error?". Our formula allows us to give sharp answers to this question, both in
the positive and negative directions. More precisely, we show that the sufficien

t conditions for Gaussian universality (or lack thereof) crucially depend on the alignment between the target weights and the means and covariances of the mixture clusters, which we precisely quantify. In the particular case of least-squares interpolation, we prove a strong universality property of the training error and show it follows a simple, closed-form expression. Finally, we apply our results to real datasets, clarifying some recent discussions in the literature about Gaussian universality of the errors in this context.
****

Certifying Ensembles: A General Certification Theory with S-Lipschitzness
Aleksandar Petrov, Francisco Eiras, Amartya Sanyal, Philip Torr, Adel Bibi
Improving and guaranteeing the robustness of deep learning models has been a topic of intense research. Ensembling, which combines several classifiers to provide a better model, has been shown to be beneficial for generalisation, uncertainty estimation, calibration, and mitigating the effects of concept drift. However, the impact of ensembling on certified robustness is less well understood. In this work, we generalise Lipschitz continuity by introducing S-Lipschitz classifiers, which we use to analyse the theoretical robustness of ensembles. Our results are precise conditions when ensembles of robust classifiers are more robust than any constituent classifier, as well as conditions when they are less robust.
****

The Power of Learned Locally Linear Models for Nonlinear Policy Optimization
Daniel Pfrommer, Max Simchowitz, Tyler Westenbroek, Nikolai Matni, Stephen Tu
A common pipeline in learning-based control is to iteratively estimate a model of system dynamics, and apply a trajectory optimization algorithm - e.g. $\mathtt{iLQR}$ - on the learned model to minimize a target cost. This paper conducts a rigorous analysis of a simplified variant of this strategy for general nonlinear systems. We analyze an algorithm which iterates between estimating local linear models of nonlinear system dynamics and performing $\mathtt{iLQR}$-like policy updates. We demonstrate that this algorithm attains sample complexity polynomial in relevant problem parameters, and, by synthesizing locally stabilizing gains, overcomes exponential dependence in problem horizon. Experimental results validate the performance of our algorithm, and compare to natural deep-learning baselines.
****

A Scalable Frank-Wolfe-Based Algorithm for the Max-Cut SDP
Chi Bach Pham, Wynita Griggs, James Saunderson
We consider the problem of solving large-scale instances of the Max-Cut semidefinite program (SDP), i.e., optimizing a linear function over $n\times n$ positive semidefinite (PSD) matrices with unit diagonal. When the cost matrix is PSD, we show how to exactly reformulate the problem as maximizing a smooth concave function over PSD matrices with unit trace. By applying the Frank-Wolfe method, we obtain a simple algorithm that is compatible with recent sampling-based techniques to solve SDPs using low memory. We demonstrate the practical performance of our method on $10^6\times 10^6$ instances of the max-cut SDP with costs having up to 5×10^6 non-zero entries. Theoretically, we show that our method solves problems with diagonally dominant costs to relative error ϵ in $O(n\epsilon^{-1})$ calls to a randomized approximate largest eigenvalue subroutine, each of which succeeds with high probability after $O(\log(n)\epsilon^{-1/2})$ matrix-vector multiplications with the cost matrix.
****

Attention-Based Recurrence for Multi-Agent Reinforcement Learning under Stochastic Partial Observability
Thomy Phan, Fabian Ritz, Philipp Altmann, Maximilian Zorn, Jonas Nüßlein, Michael Kölle, Thomas Gabor, Claudia Linnhoff-Popien
Stochastic partial observability poses a major challenge for decentralized coordination in multi-agent reinforcement learning but is largely neglected in state-of-the-art research due to a strong focus on state-based centralized training for decentralized execution (CTDE) and benchmarks that lack sufficient stochasticity like StarCraft Multi-Agent Challenge (SMAC). In this paper, we propose Attention-based Embeddings of Recurrence In multi-Agent Learning (AERIAL) to approxima

te value functions under stochastic partial observability. AERIAL replaces the t
rue state with a learned representation of multi-agent recurrence, considering m
ore accurate information about decentralized agent decisions than state-based CT
DE. We then introduce MessySMAC, a modified version of SMAC with stochastic obse
rvations and higher variance in initial states, to provide a more general and co
nfigurable benchmark regarding stochastic partial observability. We evaluate AER
IAL in Dec-Tiger as well as in a variety of SMAC and MessySMAC maps, and compare
 the results with state-based CTDE. Furthermore, we evaluate the robustness of A
ERIAL and state-based CTDE against various stochasticity configurations in Messy
SMAC.
****

HyperTuning: Toward Adapting Large Language Models without Back-propagation
Jason Phang, Yi Mao, Pengcheng He, Weizhu Chen
Fine-tuning large language models for different tasks can be costly and ineffici
ent, and even methods that reduce the number of tuned parameters still require f
ull gradient-based optimization. We propose HyperTuning, a novel approach to mod
el adaptation that uses a hypermodel to generate task-specific parameters for a
fixed downstream model. We demonstrate a simple setup for hypertuning with Hyper
T5, a T5-based hypermodel that produces soft prefixes or LoRA parameters for a f
rozen T5 model from few-shot examples. We train HyperT5 in two stages: first, hy
perpretraining with a modified conditional language modeling objective that trai
ns a hypermodel to generate parameters; second, multi-task fine-tuning (MTF) on
a large number of diverse language tasks. We evaluate HyperT5 on P3, MetaICL and
 Super-NaturalInstructions datasets, and show that it can effectively generate p
arameters for unseen tasks. Moreover, we show that using hypermodel-generated pa
rameters as initializations for further parameter-efficient fine-tuning improves
 performance. HyperTuning can thus be a flexible and efficient way to leverage l
arge language models for diverse downstream applications.
****

Linear CNNs Discover the Statistical Structure of the Dataset Using Only the Mos
t Dominant Frequencies
Hannah Pinson, Joeri Lenaerts, Vincent Ginis
We here present a stepping stone towards a deeper understanding of convolutional
 neural networks (CNNs) in the form of a theory of learning in linear CNNs. Thro
ugh analyzing the gradient descent equations, we discover that the evolution of
the network during training is determined by the interplay between the dataset s
tructure and the convolutional network structure. We show that linear CNNs disco
ver the statistical structure of the dataset with non-linear, ordered, stage-lik
e transitions, and that the speed of discovery changes depending on the relation
ship between the dataset and the convolutional network structure. Moreover, we f
ind that this interplay lies at the heart of what we call the "dominant frequenc
y bias", where linear CNNs arrive at these discoveries using only the dominant f
requencies of the different structural parts present in the dataset. We furtherm
ore provide experiments that show how our theory relates to deep, non-linear CNN
s used in practice. Our findings shed new light on the inner working of CNNs, an
d can help explain their shortcut learning and their tendency to rely on texture
 instead of shape.
****

Conformal Prediction for Federated Uncertainty Quantification Under Label Shift
Vincent Plassier, Mehdi Makni, Aleksandr Rubashevskii, Eric Moulines, Maxim Pano
v
Federated Learning (FL) is a machine learning framework where many clients colla
boratively train models while keeping the training data decentralized. Despite r
ecent advances in FL, the uncertainty quantification topic (UQ) remains partiall
y addressed. Among UQ methods, conformal prediction (CP) approaches provides dis
tribution-free guarantees under minimal assumptions. We develop a new federated
conformal prediction method based on quantile regression and take into account p
rivacy constraints. This method takes advantage of importance weighting to effec
tively address the label shift between agents and provides theoretical guarantee
s for both valid coverage of the prediction sets and differential privacy. Exten

sive experimental studies demonstrate that this method outperforms current competitors.
****
Universal Physics-Informed Neural Networks: Symbolic Differential Operator Discovery with Sparse Data
Lena Podina, Brydon Eastman, Mohammad Kohandel
In this work we perform symbolic discovery of differential operators in a situation where there is sparse experimental data. This small data regime in machine learning can be made tractable by providing our algorithms with prior information about the underlying dynamics. Physics Informed Neural Networks (PINNs) have been very successful in this regime (reconstructing entire ODE solutions using only a single point or entire PDE solutions with very few measurements of the initial condition). The Universal PINN approach (UPINN) adds a neural network that learns a representation of unknown hidden terms in the differential equation. The algorithm yields both a surrogate solution to the differential equation and a black-box representation of the hidden terms. These hidden term neural networks can then be converted into symbolic equations using symbolic regression techniques like AI Feynman. In order to achieve convergence of the neural networks, we provide our algorithms with (noisy) measurements of both the initial condition as well as (synthetic) experimental data obtained at later times. We demonstrate strong performance of UPINNs even when provided with very few measurements of noisy data in both the ODE and PDE regime.
****
Sequential Kernelized Independence Testing
Aleksandr Podkopaev, Patrick Blöbaum, Shiva Kasiviswanathan, Aaditya Ramdas
Independence testing is a classical statistical problem that has been extensively studied in the batch setting when one fixes the sample size before collecting data. However, practitioners often prefer procedures that adapt to the complexity of a problem at hand instead of setting sample size in advance. Ideally, such procedures should (a) stop earlier on easy tasks (and later on harder tasks), hence making better use of available resources, and (b) continuously monitor the data and efficiently incorporate statistical evidence after collecting new data, while controlling the false alarm rate. Classical batch tests are not tailored for streaming data: valid inference after data peeking requires correcting for multiple testing which results in low power. Following the principle of testing by betting, we design sequential kernelized independence tests that overcome such shortcomings. We exemplify our broad framework using bets inspired by kernelized dependence measures, e.g., the Hilbert-Schmidt independence criterion. Our test is also valid under non-i.i.d. time-varying settings. We demonstrate the power of our approaches on both simulated and real data.
****
Truncating Trajectories in Monte Carlo Reinforcement Learning
Riccardo Poiani, Alberto Maria Metelli, Marcello Restelli
In Reinforcement Learning (RL), an agent acts in an unknown environment to maximize the expected cumulative discounted sum of an external reward signal, i.e., the expected return. In practice, in many tasks of interest, such as policy optimization, the agent usually spends its interaction budget by collecting episodes of fixed length within a simulator (i.e., Monte Carlo simulation). However, given the discounted nature of the RL objective, this data collection strategy might not be the best option. Indeed, the rewards taken in early simulation steps weigh exponentially more than future rewards. Taking a cue from this intuition, in this paper, we design an a-priori budget allocation strategy that leads to the collection of trajectories of different lengths, i.e., truncated. The proposed approach provably minimizes the width of the confidence intervals around the empirical estimates of the expected return of a policy. After discussing the theoretical properties of our method, we make use of our trajectory truncation mechanism to extend Policy Optimization via Importance Sampling (POIS, Metelli et al., 2018) algorithm. Finally, we conduct a numerical comparison between our algorithm and POIS: the results are consistent with our theory and show that an appropriate truncation of the trajectories can succeed in improving performance.

****
Hyena Hierarchy: Towards Larger Convolutional Language Models
Michael Poli, Stefano Massaroli, Eric Nguyen, Daniel Y Fu, Tri Dao, Stephen Baccus, Yoshua Bengio, Stefano Ermon, Christopher Re
Recent advances in deep learning have relied heavily on the use of large Transformers due to their ability to learn at scale. However, the core building block of Transformers, the attention operator, exhibits quadratic cost in sequence length, limiting the amount of context accessible. Existing subquadratic methods based on low-rank and sparse approximations need to be combined with dense attention layers to match Transformers at scale, indicating a gap in capability. In this work, we propose Hyena, a subquadratic drop-in replacement for attention constructed by interleaving implicitly parametrized long convolutions and data-controlled gating. In challenging reasoning tasks on sequences of thousands to hundreds of thousands of tokens, Hyena improves accuracy by more than 50 points over operators relying on state-space models, transfer functions, and other implicit and explicit methods, matching attention-based models. We set a new state-of-the-art for dense-attention-free architectures on language modeling in standard datasets WikiText103 and The Pile, reaching Transformer quality with a 20% reduction in training compute required at sequence length 2k. Hyena operators are 2x faster than highly optimized attention at sequence length 8k, with speedups of 100x at 64k.
****
Spurious Valleys and Clustering Behavior of Neural Networks
Samuele Pollaci
Neural networks constitute a class of functions that are typically non-surjective, with high-dimensional fibers and complicated image. We prove two main results concerning the geometry of the loss landscape of a neural network. First, we provide an explicit effective bound on the sizes of the hidden layers so that the loss landscape has no spurious valleys, which guarantees the success of gradient descent methods. Second, we present a novel method for analyzing whether a given neural network architecture with monomial activation function can represent a target function of interest. The core of our analysis method is the study of a specific set of error values, and its behavior depending on different training datasets.
****
Multisample Flow Matching: Straightening Flows with Minibatch Couplings
Aram-Alexandre Pooladian, Heli Ben-Hamu, Carles Domingo-Enrich, Brandon Amos, Yaron Lipman, Ricky T. Q. Chen
Simulation-free methods for training continuous-time generative models construct probability paths that go between noise distributions and individual data samples. Recent works, such as Flow Matching, derived paths that are optimal for each data sample. However, these algorithms rely on independent data and noise samples, and do not exploit underlying structure in the data distribution for constructing probability paths. We propose Multisample Flow Matching, a more general framework that uses non-trivial couplings between data and noise samples while satisfying the correct marginal constraints. At small overhead costs, this generalization allows us to (i) reduce gradient variance during training, (ii) obtain straighter flows for the learned vector field, which allows us to generate high-quality samples using fewer function evaluations, and (iii) obtain transport maps with low cost in high dimensions, which has applications beyond generative modeling. Importantly, we do so in a completely simulation-free manner with a simple minimization objective. We show that our proposed methods improve sample consistency on downsampled ImageNet data sets, and lead to better low-cost sample generation.
****
Minimax estimation of discontinuous optimal transport maps: The semi-discrete case
Aram-Alexandre Pooladian, Vincent Divol, Jonathan Niles-Weed
We consider the problem of estimating the optimal transport map between two probability distributions, P and Q in $\mathbb{R}^d$, on the basis of i.i.d. sam

ples. All existing statistical analyses of this problem require the assumption t
hat the transport map is Lipschitz, a strong requirement that, in particular, ex
cludes any examples where the transport map is discontinuous. As a first step to
wards developing estimation procedures for discontinuous maps, we consider the i
mportant special case where the data distribution Q is a discrete measure supp
orted on a finite number of points in $\mathbb{R}^d$. We study a computationally
 efficient estimator initially proposed by (Pooladian & Niles-Weed, 2021), based
 on entropic optimal transport, and show in the semi-discrete setting that it co
nverges at the minimax-optimal rate $n^{-1/2}$, independent of dimension. Other
standard map estimation techniques both lack finite-sample guarantees in this se
tting and provably suffer from the curse of dimensionality. We confirm these res
ults in numerical experiments, and provide experiments for other settings, not c
overed by our theory, which indicate that the entropic estimator is a promising
methodology for other discontinuous transport map estimation problems.
****

Test-time Adaptation with Slot-Centric Models
Mihir Prabhudesai, Anirudh Goyal, Sujoy Paul, Sjoerd Van Steenkiste, Mehdi S. M.
 Sajjadi, Gaurav Aggarwal, Thomas Kipf, Deepak Pathak, Katerina Fragkiadaki
Current visual detectors, though impressive within their training distribution,
often fail to parse out-of-distribution scenes into their constituent entities.
Recent test-time adaptation methods use auxiliary self-supervised losses to adap
t the network parameters to each test example independently and have shown promi
sing results towards generalization outside the training distribution for the ta
sk of image classification. In our work, we find evidence that these losses are
insufficient for the task of scene decomposition, without also considering archi
tectural inductive biases. Recent slot-centric generative models attempt to deco
mpose scenes into entities in a self-supervised manner by reconstructing pixels.
 Drawing upon these two lines of work, we propose Slot-TTA, a semi-supervised sl
ot-centric scene decomposition model that at test time is adapted per scene thro
ugh gradient descent on reconstruction or cross-view synthesis objectives. We ev
aluate Slot-TTA across multiple input modalities, images or 3D point clouds, and
 show substantial out-of-distribution performance improvements against state-of-
the-art supervised feed-forward detectors, and alternative test-time adaptation
methods. Project Webpage: http://slot-tta.github.io/
****

JAWS-X: Addressing Efficiency Bottlenecks of Conformal Prediction Under Standard
 and Feedback Covariate Shift
Drew Prinster, Suchi Saria, Anqi Liu
We study the efficient estimation of predictive confidence intervals for black-b
ox predictors when the common data exchangeability (e.g., i.i.d.) assumption is
violated due to potentially feedback-induced shifts in the input data distributi
on. That is, we focus on standard and feedback covariate shift (FCS), where the
latter allows for feedback dependencies between train and test data that occur i
n many decision-making scenarios like experimental design. Whereas prior conform
al prediction methods for this problem are in general either extremely computati
onally demanding or make inefficient use of labeled data, we propose a collectio
n of methods based on the jackknife+ that achieve a practical balance of computa
tional and statistical efficiency. Theoretically, our proposed JAW-FCS method ex
tends the rigorous, finite-sample coverage guarantee of the jackknife+ to FCS. W
e moreover propose two tunable relaxations to JAW-FCS's computation that maintai
n finite-sample guarantees: one using only K leave-one-out models (JAW-KLOO)
 and a second building on K-fold cross validation+ (WCV+). Practically, we dem
onstrate that JAW-FCS and its computational relaxations outperform state-of-the-
art baselines on a variety of real-world datasets under standard and feedback co
variate shift, including for biomolecular design and active learning tasks.
****

Equivariant Polynomials for Graph Neural Networks
Omri Puny, Derek Lim, Bobak Kiani, Haggai Maron, Yaron Lipman
Graph Neural Networks (GNN) are inherently limited in their expressive power. Re
cent seminal works (Xu et al., 2019; Morris et al., 2019b) introduced the Weisfe

iler-Lehman (WL) hierarchy as a measure of expressive power. Although this hiera rchy has propelled significant advances in GNN analysis and architecture develop ments, it suffers from several significant limitations. These include a complex definition that lacks direct guidance for model improvement and a WL hierarchy t hat is too coarse to study current GNNs. This paper introduces an alternative ex pressive power hierarchy based on the ability of GNNs to calculate equivariant p olynomials of a certain degree. As a first step, we provide a full characterizat ion of all equivariant graph polynomials by introducing a concrete basis, signif icantly generalizing previous results. Each basis element corresponds to a speci fic multi-graph, and its computation over some graph data input corresponds to a tensor contraction problem. Second, we propose algorithmic tools for evaluating the expressiveness of GNNs using tensor contraction sequences, and calculate th e expressive power of popular GNNs. Finally, we enhance the expressivity of comm on GNN architectures by adding polynomial features or additional operations / ag gregations inspired by our theory. These enhanced GNNs demonstrate state-of-the- art results in experiments across multiple graph learning benchmarks.
****

Contrast with Reconstruct: Contrastive 3D Representation Learning Guided by Gene rative Pretraining
Zekun Qi, Runpei Dong, Guofan Fan, Zheng Ge, Xiangyu Zhang, Kaisheng Ma, Li Yi
Mainstream 3D representation learning approaches are built upon contrastive or g enerative modeling pretext tasks, where great improvements in performance on var ious downstream tasks have been achieved. However, we find these two paradigms h ave different characteristics: (i) contrastive models are data-hungry that suffe r from a representation over-fitting issue; (ii) generative models have a data f illing issue that shows inferior data scaling capacity compared to contrastive m odels. This motivates us to learn 3D representations by sharing the merits of bo th paradigms, which is non-trivial due to the pattern difference between the two paradigms. In this paper, we propose contrast with reconstruct (ReCon) that uni fies these two paradigms. ReCon is trained to learn from both generative modelin g teachers and cross-modal contrastive teachers through ensemble distillation, w here the generative student is used to guide the contrastive student. An encoder -decoder style ReCon-block is proposed that transfers knowledge through cross at tention with stop-gradient, which avoids pretraining over-fitting and pattern di fference issues. ReCon achieves a new state-of-the-art in 3D representation lear ning, e.g., 91.26% accuracy on ScanObjectNN. Codes have been released at https:/ /github.com/qizekun/ReCon.
****

An Effective Meaningful Way to Evaluate Survival Models
Shi-Ang Qi, Neeraj Kumar, Mahtab Farrokh, Weijie Sun, Li-Hao Kuan, Rajesh Rangan ath, Ricardo Henao, Russell Greiner
One straightforward metric to evaluate a survival prediction model is based on t he Mean Absolute Error (MAE) – the average of the absolute difference between th e time predicted by the model and the true event time, over all subjects. Unfort unately, this is challenging because, in practice, the test set includes (right) censored individuals, meaning we do not know when a censored individual actuall y experienced the event. In this paper, we explore various metrics to estimate M AE for survival datasets that include (many) censored individuals. Moreover, we introduce a novel and effective approach for generating realistic semi-synthetic survival datasets to facilitate the evaluation of metrics. Our findings, based on the analysis of the semi-synthetic datasets, reveal that our proposed metric (MAE using pseudo-observations) is able to rank models accurately based on their performance, and often closely matches the true MAE – in particular, is better than several alternative methods.
****

Coarse-to-Fine: a Hierarchical Diffusion Model for Molecule Generation in 3D
Bo Qiang, Yuxuan Song, Minkai Xu, Jingjing Gong, Bowen Gao, Hao Zhou, Wei-Ying M a, Yanyan Lan
Generating desirable molecular structures in 3D is a fundamental problem for dru g discovery. Despite the considerable progress we have achieved, existing method

s usually generate molecules in atom resolution and ignore intrinsic local structures such as rings, which leads to poor quality in generated structures, especially when generating large molecules. Fragment-based molecule generation is a promising strategy, however, it is nontrivial to be adapted for 3D non-autoregressive generations because of the combinational optimization problems. In this paper, we utilize a coarse-to-fine strategy to tackle this problem, in which a Hierarchical Diffusion-based model (i.e. HierDiff) is proposed to preserve the validity of local segments without relying on autoregressive modeling. Specifically, HierDiff first generates coarse-grained molecule geometries via an equivariant diffusion process, where each coarse-grained node reflects a fragment in a molecule. Then the coarse-grained nodes are decoded into fine-grained fragments by a message-passing process and a newly designed iterative refined sampling module. Lastly, the fine-grained fragments are then assembled to derive a complete atomic molecular structure. Extensive experiments demonstrate that HierDiff consistently improves the quality of molecule generation over existing methods.
****

Collaborative Causal Inference with Fair Incentives
Rui Qiao, Xinyi Xu, Bryan Kian Hsiang Low
Collaborative causal inference (CCI) aims to improve the estimation of the causal effect of treatment variables by utilizing data aggregated from multiple self-interested parties. Since their source data are valuable proprietary assets that can be costly or tedious to obtain, every party has to be incentivized to be willing to contribute to the collaboration, such as with a guaranteed fair and sufficiently valuable reward (than performing causal inference on its own). This paper presents a reward scheme designed using the unique statistical properties that are required by causal inference to guarantee certain desirable incentive criteria (e.g., fairness, benefit) for the parties based on their contributions. To achieve this, we propose a data valuation function to value parties' data for CCI based on the distributional closeness of its resulting treatment effect estimate to that utilizing the aggregated data from all parties. Then, we show how to value the parties' rewards fairly based on a modified variant of the Shapley value arising from our proposed data valuation for CCI. Finally, the Shapley fair rewards to the parties are realized in the form of improved, stochastically perturbed treatment effect estimates. We empirically demonstrate the effectiveness of our reward scheme using simulated and real-world datasets.
****

FREDIS: A Fusion Framework of Refinement and Disambiguation for Unreliable Partial Label Learning
Congyu Qiao, Ning Xu, Jiaqi Lv, Yi Ren, Xin Geng
To reduce the difficulty of annotation, partial label learning (PLL) has been widely studied, where each example is ambiguously annotated with a set of candidate labels instead of the exact correct label. PLL assumes that the candidate label set contains the correct label, which induces disambiguation, i.e., identification of the correct label in the candidate label set, adopted in most PLL methods. However, this assumption is impractical as no one could guarantee the existence of the correct label in the candidate label set under real-world scenarios. Therefore, Unreliable Partial Label Learning (UPLL) is investigated where the correct label of each example may not exist in the candidate label set. In this paper, we propose a fusion framework of refinement and disambiguation named FREDIS to handle the UPLL problem. Specifically, with theoretical guarantees, not only does disambiguation move incorrect labels from candidate labels to non-candidate labels but also refinement, an opposite procedure, moves correct labels from non-candidate labels to candidate labels. Besides, we prove that the classifier trained by our framework could eventually approximate the Bayes optimal classifier. Extensive experiments on widely used benchmark datasets validate the effectiveness of our proposed framework.
****

Nugget: Neural Agglomerative Embeddings of Text
Guanghui Qin, Benjamin Van Durme
Embedding text sequences is a widespread requirement in modern language understa

nding. Existing approaches focus largely on constant-size representations. This is problematic, as the amount of information contained in text often varies with the length of the input. We propose a solution called Nugget, which encodes language into a representation based on a dynamically selected subset of input tokens. These nuggets are learned through tasks like autoencoding and machine translation, and intuitively segment language into meaningful units. We demonstrate Nugget outperforms related approaches in tasks involving semantic comparison. Finally, we illustrate these compact units allow for expanding the contextual window of a language model (LM), suggesting new future LMs that can condition on significantly larger amounts of content.

****

BiBench: Benchmarking and Analyzing Network Binarization

Haotong Qin, Mingyuan Zhang, Yifu Ding, Aoyu Li, Zhongang Cai, Ziwei Liu, Fisher Yu, Xianglong Liu

Network binarization emerges as one of the most promising compression approaches offering extraordinary computation and memory savings by minimizing the bit-width. However, recent research has shown that applying existing binarization algorithms to diverse tasks, architectures, and hardware in realistic scenarios is still not straightforward. Common challenges of binarization, such as accuracy degradation and efficiency limitation, suggest that its attributes are not fully understood. To close this gap, we present BiBench, a rigorously designed benchmark with in-depth analysis for network binarization. We first carefully scrutinize the requirements of binarization in the actual production and define evaluation tracks and metrics for a comprehensive and fair investigation. Then, we evaluate and analyze a series of milestone binarization algorithms that function at the operator level and with extensive influence. Our benchmark reveals that 1) the binarized operator has a crucial impact on the performance and deployability of binarized networks; 2) the accuracy of binarization varies significantly across different learning tasks and neural architectures; 3) binarization has demonstrated promising efficiency potential on edge devices despite the limited hardware support. The results and analysis also lead to a promising paradigm for accurate and efficient binarization. We believe that BiBench will contribute to the broader adoption of binarization and serve as a foundation for future research. The code for our BiBench is released https://github.com/htqin/BiBench .

****

Not All Semantics are Created Equal: Contrastive Self-supervised Learning with Automatic Temperature Individualization

Zi-Hao Qiu, Quanqi Hu, Zhuoning Yuan, Denny Zhou, Lijun Zhang, Tianbao Yang

In this paper, we aim to optimize a contrastive loss with individualized temperatures in a principled manner. The common practice of using a global temperature parameter τ ignores the fact that “not all semantics are created equal", meaning that different anchor data may have different numbers of samples with similar semantics, especially when data exhibits long-tails. First, we propose a new robust contrastive loss inspired by distributionally robust optimization (DRO), providing us an intuition about the effect of τ and a mechanism for automatic temperature individualization. Then, we propose an efficient stochastic algorithm for optimizing the robust contrastive loss with a provable convergence guarantee without using large mini-batch sizes. Theoretical and experimental results show that our algorithm automatically learns a suitable τ for each sample. Specifically, samples with frequent semantics use large temperatures to keep local semantic structures, while samples with rare semantics use small temperatures to induce more separable features. Our method not only outperforms prior strong baselines (e.g., SimCLR, CLIP) on unimodal and bimodal tasks with larger improvements on imbalanced data but also is less sensitive to hyper-parameters. To our best knowledge, this is the first methodical approach to optimizing a contrastive loss with individualized temperatures. Our proposed algorithms are implemented in the LibAUC library at https://libauc.org.

****

Shortest Edit Path Crossover: A Theory-driven Solution to the Permutation Problem in Evolutionary Neural Architecture Search

Xin Qiu, Risto Miikkulainen
Population-based search has recently emerged as a possible alternative to Reinfo rcement Learning (RL) for black-box neural architecture search (NAS). It perform s well in practice even though it is not theoretically well understood. In parti cular, whereas traditional population-based search methods such as evolutionary algorithms (EAs) draw much power from crossover operations, it is difficult to t ake advantage of them in NAS. The main obstacle is believed to be the permutatio n problem: The mapping between genotype and phenotype in traditional graph repre sentations is many-to-one, leading to a disruptive effect of standard crossover. This paper presents the first theoretical analysis of the behaviors of mutation , crossover and RL in black-box NAS, and proposes a new crossover operator based on the shortest edit path (SEP) in graph space. The SEP crossover is shown theo retically to overcome the permutation problem, and as a result, have a better ex pected improvement compared to mutation, standard crossover and RL. Further, it empirically outperform these other methods on state-of-the-art NAS benchmarks. T he SEP crossover therefore allows taking full advantage of population-based sear ch in NAS, and the underlying theory can serve as a foundation for deeper unders tanding of black-box NAS methods in general.
****

Simple and Fast Group Robustness by Automatic Feature Reweighting
Shikai Qiu, Andres Potapczynski, Pavel Izmailov, Andrew Gordon Wilson
A major challenge to out-of-distribution generalization is reliance on spurious features — patterns that are predictive of the class label in the training data distribution, but not causally related to the target. Standard methods for reduc ing the reliance on spurious features typically assume that we know what the spu rious feature is, which is rarely true in the real world. Methods that attempt t o alleviate this limitation are complex, hard to tune, and lead to a significant computational overhead compared to standard training. In this paper, we propose Automatic Feature Reweighting (AFR), an extremely simple and fast method for up dating the model to reduce the reliance on spurious features. AFR retrains the l ast layer of a standard ERM-trained base model with a weighted loss that emphasi zes the examples where the ERM model predicts poorly, automatically upweighting the minority group without group labels. With this simple procedure, we improve upon the best reported results among competing methods trained without spurious attributes on several vision and natural language classification benchmarks, usi ng only a fraction of their compute.
****

DRCFS: Doubly Robust Causal Feature Selection
Francesco Quinzan, Ashkan Soleymani, Patrick Jaillet, Cristian R. Rojas, Stefan Bauer
Knowing the features of a complex system that are highly relevant to a particula r target variable is of fundamental interest in many areas of science. Existing approaches are often limited to linear settings, sometimes lack guarantees, and in most cases, do not scale to the problem at hand, in particular to images. We propose DRCFS, a doubly robust feature selection method for identifying the caus al features even in nonlinear and high dimensional settings. We provide theoreti cal guarantees, illustrate necessary conditions for our assumptions, and perform extensive experiments across a wide range of simulated and semi-synthetic datas ets. DRCFS significantly outperforms existing state-of-the-art methods, selectin g robust features even in challenging highly non-linear and high-dimensional pro blems.
****

Robust Speech Recognition via Large-Scale Weak Supervision
Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, Ilya Sut skever
We study the capabilities of speech processing systems trained simply to predict large amounts of transcripts of audio on the internet. When scaled to 680,000 h ours of multilingual and multitask supervision, the resulting models generalize well to standard benchmarks and are often competitive with prior fully supervise d results without the need for any dataset specific fine-tuning. When compared t

o humans, the models approach their accuracy and robustness. We are releasing mo
dels and inference code to serve as a foundation for further work on robust spee
ch processing.
****

Shiftable Context: Addressing Training-Inference Context Mismatch in Simultaneou
s Speech Translation

Matthew Raffel, Drew Penney, Lizhong Chen

Transformer models using segment-based processing have been an effective archite
cture for simultaneous speech translation. However, such models create a context
 mismatch between training and inference environments, hindering potential trans
lation accuracy. We solve this issue by proposing Shiftable Context, a simple ye
t effective scheme to ensure that consistent segment and context sizes are maint
ained throughout training and inference, even with the presence of partially fil
led segments due to the streaming nature of simultaneous translation. Shiftable
Context is also broadly applicable to segment-based transformers for streaming t
asks. Our experiments on the English-German, English-French, and English-Spanish
 language pairs from the MUST-C dataset demonstrate that when applied to the Aug
mented Memory Transformer, a state-of-the-art model for simultaneous speech tran
slation, the proposed scheme achieves an average increase of 2.09, 1.83, and 1.9
5 BLEU scores across each wait-k value for the three language pairs, respectivel
y, with a minimal impact on computation-aware Average Lagging.
****

Sequential Multi-Dimensional Self-Supervised Learning for Clinical Time Series

Aniruddh Raghu, Payal Chandak, Ridwan Alam, John Guttag, Collin Stultz

Self-supervised learning (SSL) for clinical time series data has received signif
icant attention in recent literature, since these data are highly rich and provi
de important information about a patient’s physiological state. However, most ex
isting SSL methods for clinical time series are limited in that they are designe
d for unimodal time series, such as a sequence of structured features (e.g., lab
 values and vitals signs) or an individual high-dimensional physiological signal
 (e.g., an electrocardiogram). These existing methods cannot be readily extended
 to model time series that exhibit multimodality, with structured features and h
igh-dimensional data being recorded at each timestep in the sequence. In this wo
rk, we address this gap and propose a new SSL method — Sequential Multi-Dimensio
nal SSL — where a SSL loss is applied both at the level of the entire sequence a
nd at the level of the individual high-dimensional data points in the sequence i
n order to better capture information at both scales. Our strategy is agnostic t
o the specific form of loss function used at each level – it can be contrastive,
 as in SimCLR, or non-contrastive, as in VICReg. We evaluate our method on two r
eal-world clinical datasets, where the time series contains sequences of (1) hig
h-frequency electrocardiograms and (2) structured data from lab values and vital
s signs. Our experimental results indicate that pre-training with our method and
 then fine-tuning on downstream tasks improves performance over baselines on bot
h datasets, and in several settings, can lead to improvements across different s
elf-supervised loss functions.
****

Recovery Bounds on Class-Based Optimal Transport: A Sum-of-Norms Regularization
Framework

Arman Rahbar, Ashkan Panahi, Morteza Haghir Chehreghani, Devdatt Dubhashi, Hamid
 Krim

We develop a novel theoretical framework for understating Optimal Transport (OT)
 schemes respecting a class structure. For this purpose, we propose a convex OT
program with a sum-of-norms regularization term, which provably recovers the und
erlying class structure under geometric assumptions. Furthermore, we derive an a
ccelerated proximal algorithm with a closed-form projection and proximal operato
r scheme, thereby affording a more scalable algorithm for computing optimal tran
sport plans. We provide a novel argument for the uniqueness of the optimum even
in the absence of strong convexity. Our experiments show that the new regularize
r not only results in a better preservation of the class structure in the data b
ut also yields additional robustness to the data geometry, compared to previous

regularizers.
****
Algorithmic Stability of Heavy-Tailed SGD with General Loss Functions
Anant Raj, Lingjiong Zhu, Mert Gurbuzbalaban, Umut Simsekli
Heavy-tail phenomena in stochastic gradient descent (SGD) have been reported in several empirical studies. Experimental evidence in previous works suggests a strong interplay between the heaviness of the tails and generalization behavior of SGD. To address this empirical phenomena theoretically, several works have made strong topological and statistical assumptions to link the generalization error to heavy tails. Very recently, new generalization bounds have been proven, indicating a non-monotonic relationship between the generalization error and heavy tails, which is more pertinent to the reported empirical observations. While these bounds do not require additional topological assumptions given that SGD can be modeled using a heavy-tailed stochastic differential equation (SDE), they can only apply to simple quadratic problems. In this paper, we build on this line of research and develop generalization bounds for a more general class of objective functions, which includes non-convex functions as well. Our approach is based on developing Wasserstein stability bounds for heavy-tailed SDEs and their discretizations, which we then convert to generalization bounds. Our results do not require any nontrivial assumptions; yet, they shed more light to the empirical observations, thanks to the generality of the loss functions.
****
Mastering the Unsupervised Reinforcement Learning Benchmark from Pixels
Sai Rajeswar, Pietro Mazzaglia, Tim Verbelen, Alexandre Piché, Bart Dhoedt, Aaron Courville, Alexandre Lacoste
Controlling artificial agents from visual sensory data is an arduous task. Reinforcement learning (RL) algorithms can succeed but require large amounts of interactions between the agent and the environment. To alleviate the issue, unsupervised RL proposes to employ self-supervised interaction and learning, for adapting faster to future tasks. Yet, as shown in the Unsupervised RL Benchmark (URLB; Laskin et al. 2021), whether current unsupervised strategies can improve generalization capabilities is still unclear, especially in visual control settings. In this work, we study the URLB and propose a new method to solve it, using unsupervised model-based RL, for pre-training the agent, and a task-aware fine-tuning strategy combined with a new proposed hybrid planner, Dyna-MPC, to adapt the agent for downstream tasks. On URLB, our method obtains 93.59% overall normalized performance, surpassing previous baselines by a staggering margin. The approach is empirically evaluated through a large-scale empirical study, which we use to validate our design choices and analyze our models. We also show robust performance on the Real-Word RL benchmark, hinting at resiliency to environment perturbations during adaptation. Project website: https://masteringurlb.github.io/
****
SpotEM: Efficient Video Search for Episodic Memory
Santhosh Kumar Ramakrishnan, Ziad Al-Halah, Kristen Grauman
The goal in episodic memory (EM) is to search a long egocentric video to answer a natural language query (e.g., “where did I leave my purse?”). Existing EM methods exhaustively extract expensive fixed-length clip features to look everywhere in the video for the answer, which is infeasible for long wearable-camera videos that span hours or even days. We propose SpotEM, an approach to achieve efficiency for a given EM method while maintaining good accuracy. SpotEM consists of three key ideas: 1) a novel clip selector that learns to identify promising video regions to search conditioned on the language query; 2) a set of low-cost semantic indexing features that capture the context of rooms, objects, and interactions that suggest where to look; and 3) distillation losses that address the optimization issues arising from end-to-end joint training of the clip selector and EM model. Our experiments on 200+ hours of video from the Ego4D EM Natural Language Queries benchmark and three different EM models demonstrate the effectiveness of our approach: computing only 10% – 25% of the clip features, we preserve 84% – 97% of the original EM model’s accuracy. Project page: https://vision.cs.utexas.edu/projects/spotem

****
How much does Initialization Affect Generalization?
Sameera Ramasinghe, Lachlan Ewen Macdonald, Moshiur Farazi, Hemanth Saratchandra n, Simon Lucey
Characterizing the remarkable generalization properties of over-parameterized ne ural networks remains an open problem. A growing body of recent literature shows that the bias of stochastic gradient descent (SGD) and architecture choice impl icitly leads to better generalization. In this paper, we show on the contrary th at, independently of architecture, SGD can itself be the cause of poor generaliz ation if one does not ensure good initialization. Specifically, we prove that an y differentiably parameterized model, trained under gradient flow, obeys a weak spectral bias law which states that sufficiently high frequencies train arbitrar ily slowly. This implies that very high frequencies present at initialization wi ll remain after training, and hamper generalization. Further, we empirically tes t the developed theoretical insights using practical, deep networks. Finally, we contrast our framework with that supplied by the flat-minima conjecture and sho w that Fourier analysis grants a more reliable framework for understanding the g eneralization of neural networks.
****
Model Ratatouille: Recycling Diverse Models for Out-of-Distribution Generalizati on
Alexandre Rame, Kartik Ahuja, Jianyu Zhang, Matthieu Cord, Leon Bottou, David Lo pez-Paz
Foundation models are redefining how AI systems are built. Practitioners now fol low a standard procedure to build their machine learning solutions: from a pre-t rained foundation model, they fine-tune the weights on the target task of intere st. So, the Internet is swarmed by a handful of foundation models fine-tuned on many diverse tasks: these individual fine-tunings exist in isolation without ben efiting from each other. In our opinion, this is a missed opportunity, as these specialized models contain rich and diverse features. In this paper, we thus pro pose model ratatouille, a new strategy to recycle the multiple fine-tunings of t he same foundation model on diverse auxiliary tasks. Specifically, we repurpose these auxiliary weights as initializations for multiple parallel fine-tunings on the target task; then, we average all fine-tuned weights to obtain the final mo del. This recycling strategy aims at maximizing the diversity in weights by leve raging the diversity in auxiliary tasks. Empirically, it improves the state of t he art on the reference DomainBed benchmark for out-of-distribution generalizati on. Looking forward, this work contributes to the emerging paradigm of updatable machine learning where, akin to open-source software development, the community collaborates to reliably update machine learning models.
****
A Picture of the Space of Typical Learnable Tasks
Rahul Ramesh, Jialin Mao, Itay Griniasty, Rubing Yang, Han Kheng Teoh, Mark Tran strum, James Sethna, Pratik Chaudhari
We develop information geometric techniques to understand the representations le arned by deep networks when they are trained on different tasks using supervised , meta-, semi-supervised and contrastive learning. We shed light on the followin g phenomena that relate to the structure of the space of tasks: (1) the manifold of probabilistic models trained on different tasks using different representati on learning methods is effectively low-dimensional; (2) supervised learning on o ne task results in a surprising amount of progress even on seemingly dissimilar tasks; progress on other tasks is larger if the training task has diverse classe s; (3) the structure of the space of tasks indicated by our analysis is consiste nt with parts of the Wordnet phylogenetic tree; (4) episodic meta-learning algor ithms and supervised learning traverse different trajectories during training bu t they fit similar models eventually; (5) contrastive and semi-supervised learni ng methods traverse trajectories similar to those of supervised learning. We use classification tasks constructed from the CIFAR-10 and Imagenet datasets to stu dy these phenomena. Code is available at https://github.com/grasp-lyrl/picture_o f_space_of_tasks.

****
Policy Regularization with Dataset Constraint for Offline Reinforcement Learning
Yuhang Ran, Yi-Chen Li, Fuxiang Zhang, Zongzhang Zhang, Yang Yu
We consider the problem of learning the best possible policy from a fixed dataset, known as offline Reinforcement Learning (RL). A common taxonomy of existing offline RL works is policy regularization, which typically constrains the learned policy by distribution or support of the behavior policy. However, distribution and support constraints are overly conservative since they both force the policy to choose similar actions as the behavior policy when considering particular states. It will limit the learned policy's performance, especially when the behavior policy is sub-optimal. In this paper, we find that regularizing the policy towards the nearest state-action pair can be more effective and thus propose Policy Regularization with Dataset Constraint (PRDC). When updating the policy in a given state, PRDC searches the entire dataset for the nearest state-action sample and then restricts the policy with the action of this sample. Unlike previous works, PRDC can guide the policy with proper behaviors from the dataset, allowing it to choose actions that do not appear in the dataset along with the given state. It is a softer constraint but still keeps enough conservatism from out-of-distribution actions. Empirical evidence and theoretical analysis show that PRDC can alleviate offline RL's fundamentally challenging value overestimation issue with a bounded performance gap. Moreover, on a set of locomotion and navigation tasks, PRDC achieves state-of-the-art performance compared with existing methods. Code is available at https://github.com/LAMDA-RL/PRDC
****
SpENCNN: Orchestrating Encoding and Sparsity for Fast Homomorphically Encrypted Neural Network Inference
Ran Ran, Xinwei Luo, Wei Wang, Tao Liu, Gang Quan, Xiaolin Xu, Caiwen Ding, Wujie Wen
Homomorphic Encryption (HE) is a promising technology to protect clients' data privacy for Machine Learning as a Service (MLaaS) on public clouds. However, HE operations can be orders of magnitude slower than their counterparts for plaintexts and thus result in prohibitively high inference latency, seriously hindering the practicality of HE. In this paper, we propose a HE-based fast neural network (NN) inference framework–SpENCNN built upon the co-design of HE operation-aware model sparsity and the single-instruction-multiple-data (SIMD)-friendly data packing, to improve NN inference latency. In particular, we first develop an encryption-aware HE-group convolution technique that can partition channels among different groups based on the data size and ciphertext size, and then encode them into the same ciphertext by novel group-interleaved encoding, so as to dramatically reduce the number of bottlenecked operations in HE convolution. We further tailor a HE-friendly sub-block weight pruning to reduce the costly HE-based convolution operation. Our experiments show that SpENCNN can achieve overall speedups of 8.37$\times$, 12.11$\times$, 19.26$\times$, and 1.87$\times$ for LeNet, VGG-5, HEFNet, and ResNet-20 respectively, with negligible accuracy loss. Our code is publicly available at https://github.com/ranran0523/SPECNN.
****
Feature learning in deep classifiers through Intermediate Neural Collapse
Akshay Rangamani, Marius Lindegaard, Tomer Galanti, Tomaso A Poggio
In this paper, we conduct an empirical study of the feature learning process in deep classifiers. Recent research has identified a training phenomenon called Neural Collapse (NC), in which the top-layer feature embeddings of samples from the same class tend to concentrate around their means, and the top layer's weights align with those features. Our study aims to investigate if these properties extend to intermediate layers. We empirically study the evolution of the covariance and mean of representations across different layers and show that as we move deeper into a trained neural network, the within-class covariance decreases relative to the between-class covariance. Additionally, we find that in the top layers, where the between-class covariance is dominant, the subspace spanned by the class means aligns with the subspace spanned by the most significant singular vector components of the weight matrix in the corresponding layer. Finally, we disc

uss the relationship between NC and Associative Memories (Willshaw et. al. 1969)
.
****

The Unintended Consequences of Discount Regularization: Improving Regularization
 in Certainty Equivalence Reinforcement Learning

Sarah Rathnam, Sonali Parbhoo, Weiwei Pan, Susan Murphy, Finale Doshi-Velez

Discount regularization, using a shorter planning horizon when calculating the o
ptimal policy, is a popular choice to restrict planning to a less complex set of
 policies when estimating an MDP from sparse or noisy data (Jiang et al., 2015).
 It is commonly understood that discount regularization functions by de-emphasiz
ing or ignoring delayed effects. In this paper, we reveal an alternate view of d
iscount regularization that exposes unintended consequences. We demonstrate that
 planning under a lower discount factor produces an identical optimal policy to
planning using any prior on the transition matrix that has the same distribution
 for all states and actions. In fact, it functions like a prior with stronger re
gularization on state-action pairs with more transition data. This leads to poor
 performance when the transition matrix is estimated from data sets with uneven
amounts of data across state-action pairs. Our equivalence theorem leads to an e
xplicit formula to set regularization parameters locally for individual state-ac
tion pairs rather than globally. We demonstrate the failures of discount regular
ization and how we remedy them using our state-action-specific method across sim
ple empirical examples as well as a medical cancer simulator.
****

Beam Tree Recursive Cells

Jishnu Ray Chowdhury, Cornelia Caragea

We propose Beam Tree Recursive Cell (BT-Cell) - a backpropagation-friendly frame
work to extend Recursive Neural Networks (RvNNs) with beam search for latent str
ucture induction. We further extend this framework by proposing a relaxation of
the hard top-k operators in beam search for better propagation of gradient sig
nals. We evaluate our proposed models in different out-of-distribution splits in
 both synthetic and realistic data. Our experiments show that BT-Cell achieves n
ear-perfect performance on several challenging structure-sensitive synthetic tas
ks like ListOps and logical inference while maintaining comparable performance i
n realistic data against other RvNN-based models. Additionally, we identify a pr
eviously unknown failure case for neural models in generalization to unseen numb
er of arguments in ListOps. The code is available at: https://github.com/JRC1995
/BeamTreeRecursiveCells.
****

Monotonic Location Attention for Length Generalization

Jishnu Ray Chowdhury, Cornelia Caragea

We explore different ways to utilize position-based cross-attention in seq2seq n
etworks to enable length generalization in algorithmic tasks. We show that a sim
ple approach of interpolating the original and reversed encoded representations
combined with relative attention allows near-perfect length generalization for b
oth forward and reverse lookup tasks or copy tasks that had been generally hard
to tackle. We also devise harder diagnostic tasks where the relative distance of
 the ideal attention position varies with timestep. In such settings, the simple
 interpolation trick with relative attention is not sufficient. We introduce nov
el variants of location attention building on top of Dubois et al. (2020) to add
ress the new diagnostic tasks. We also show the benefits of our approaches for l
ength generalization in SCAN (Lake & Baroni, 2018) and CFQ (Keysers et al.,2020)
. Our code is available on GitHub.
****

Automated Search for Conjectures on Mathematical Constants using Analysis of Int
eger Sequences

Ofir Razon, Yoav Harris, Shahar Gottlieb, Dan Carmon, Ofir David, Ido Kaminer

The discovery of formulas involving mathematical constants such as π and e
 had a great impact on various fields of science and mathematics. However, such
discoveries have remained scarce, relying on the intuition of mathematicians suc
h as Ramanujan and Gauss. Recent efforts to automate such discoveries, such as t

he Ramanujan Machine project, relied solely on exhaustive search and remain limi
ted by the space of options that can be covered. Here we propose a fundamentally
 different method to search for conjectures on mathematical constants: through a
nalysis of integer sequences. We introduce the Enumerated Signed-continued-fract
ion Massey Approve (ESMA) algorithm, which builds on the Berlekamp-Massey algori
thm to identify patterns in integer sequences that represent mathematical consta
nts. ESMA has found various known formulas and new conjectures for $e, e^2, \tan
(1)$, and ratios of values of Bessel functions, many of which provide faster num
erical convergence than their corresponding simple continued fractions forms. We
 also characterize the space of constants that ESMA can catch and quantify its a
lgorithmic advantage in certain scenarios. Altogether, this work continues the d
evelopment toward algorithm-augmented mathematical intuition, to help accelerate
 mathematical research.
****

Neural networks trained with SGD learn distributions of increasing complexity
Maria Refinetti, Alessandro Ingrosso, Sebastian Goldt
The uncanny ability of over-parameterised neural networks to generalise well has
 been explained using various "simplicity biases". These theories postulate that
 neural networks avoid overfitting by first fitting simple, linear classifiers b
efore learning more complex, non-linear functions. Meanwhile, data structure is
also recognised as a key ingredient for good generalisation, yet its role in sim
plicity biases is not yet understood. Here, we show that neural networks trained
 using stochastic gradient descent initially classify their inputs using lower-o
rder input statistics, like mean and covariance, and exploit higher-order statis
tics only later during training. We first demonstrate this distributional simpli
city bias (DSB) in a solvable model of a single neuron trained on synthetic data
. We then demonstrate DSB empirically in a range of deep convolutional networks
and visual transformers trained on CIFAR10, and show that it even holds in netwo
rks pre-trained on ImageNet. We discuss the relation of DSB to other simplicity
biases and consider its implications for the principle of Gaussian universality
in learning.
****

Simplex Random Features
Isaac Reid, Krzysztof Marcin Choromanski, Valerii Likhosherstov, Adrian Weller
We present Simplex Random Features (SimRFs), a new random feature (RF) mechanism
 for unbiased approximation of the softmax and Gaussian kernels by geometrical c
orrelation of random projection vectors. We prove that SimRFs provide the smalle
st possible mean square error (MSE) on unbiased estimates of these kernels among
 the class of weight-independent geometrically-coupled positive random feature (
PRF) mechanisms, substantially outperforming the previously most accurate Orthog
onal Random Features (ORFs) at no observable extra cost. We present a more compu
tationally expensive SimRFs+ variant, which we prove is asymptotically optimal i
n the broader family of weight-dependent geometrical coupling schemes (which per
mit correlations between random vector directions and norms). In extensive empir
ical studies, we show consistent gains provided by SimRFs in settings including
pointwise kernel estimation, nonparametric classification and scalable Transform
ers.
****

Bayesian Neural Networks Avoid Encoding Complex and Perturbation-Sensitive Conce
pts
Qihan Ren, Huiqi Deng, Yunuo Chen, Siyu Lou, Quanshi Zhang
In this paper, we focus on mean-field variational Bayesian Neural Networks (BNNs
) and explore the representation capacity of such BNNs by investigating which ty
pes of concepts are less likely to be encoded by the BNN. It has been observed a
nd studied that a relatively small set of interactive concepts usually emerge in
 the knowledge representation of a sufficiently-trained neural network, and such
 concepts can faithfully explain the network output. Based on this, our study pr
oves that compared to standard deep neural networks (DNNs), it is less likely fo
r BNNs to encode complex concepts. Experiments verify our theoretical proofs. No
te that the tendency to encode less complex concepts does not necessarily imply

weak representation power, considering that complex concepts exhibit low generalization power and high adversarial vulnerability. The code is available at https://github.com/sjtu-xai-lab/BNN-concepts.

****

Escaping saddle points in zeroth-order optimization: the power of two-point estimators

Zhaolin Ren, Yujie Tang, Na Li

Two-point zeroth order methods are important in many applications of zeroth-order optimization arising in robotics, wind farms, power systems, online optimization, and adversarial robustness to black-box attacks in deep neural networks, where the problem can be high-dimensional and/or time-varying. Furthermore, such problems may be nonconvex and contain saddle points. While existing works have shown that zeroth-order methods utilizing $\Omega(d)$ function valuations per iteration (with d denoting the problem dimension) can escape saddle points efficiently, it remains an open question if zeroth-order methods based on two-point estimators can escape saddle points. In this paper, we show that by adding an appropriate isotropic perturbation at each iteration, a zeroth-order algorithm based on $2m$ (for any $1 \leq m \leq d$) function evaluations per iteration can not only find ϵ-second order stationary points polynomially fast, but do so using only $\tilde{O}(\frac{d}{m\epsilon^{2}\bar{\psi}})$ function evaluations, where $\bar{\psi} \geq \tilde{\Omega}(\sqrt{\epsilon})$ is a parameter capturing the extent to which the function of interest exhibits the strict saddle property.

****

Dimension-independent Certified Neural Network Watermarks via Mollifier Smoothing

Jiaxiang Ren, Yang Zhou, Jiayin Jin, Lingjuan Lyu, Da Yan

Certified_Watermarks is the first to provide a watermark certificate against l_2-norm watermark removal attacks, by leveraging the randomized smoothing techniques for certified robustness to adversarial attacks. However, the randomized smoothing techniques suffer from hardness of certified robustness in high-dimensional space against l_p-norm attacks for large p ($p>2$). The certified watermark method based on the randomized smoothing is no exception, i.e., fails to provide meaningful certificates in high-dimensional space against the l_p-norm watermark removal attacks ($p>2$). By leveraging mollifier theory, this paper proposes a mollifier smoothing method with dimension-independent certified radius of our proposed smooth classifier, for conducting the certified watermark problem against the l_p-norm watermark removal attacks ($1 \leq p \leq \infty$) for high parameter dimension d. Based on partial differential equation (PDE) theory, an approximation of mollifier smoothing is developed to alleviate the inefficiency of sampling and prediction in the randomized smoothing as well as numerical integration in the mollifier smoothing, while maintaining the certified watermark against the l_p-norm watermark removal attacks ($1 \leq p \leq \infty$).

****

Feature Programming for Multivariate Time Series Prediction

Alex Daniel Reneau, Jerry Yao-Chieh Hu, Ammar Gilani, Han Liu

We introduce the concept of programmable feature engineering for time series modeling and propose a feature programming framework. This framework generates large amounts of predictive features for noisy multivariate time series while allowing users to incorporate their inductive bias with minimal effort. The key motivation of our framework is to view any multivariate time series as a cumulative sum of fine-grained trajectory increments, with each increment governed by a novel spin-gas dynamical Ising model. This fine-grained perspective motivates the development of a parsimonious set of operators that summarize multivariate time series in an abstract fashion, serving as the foundation for large-scale automated feature engineering. Numerically, we validate the efficacy of our method on several synthetic and real-world noisy time series datasets.

****

Run-off Election: Improved Provable Defense against Data Poisoning Attacks

Keivan Rezaei, Kiarash Banihashem, Atoosa Chegini, Soheil Feizi

In data poisoning attacks, an adversary tries to change a model’s prediction by adding, modifying, or removing samples in the training data. Recently, ensemble-based approaches for obtaining provable defenses against data poisoning have bee n proposed where predictions are done by taking a majority vote across multiple base models. In this work, we show that merely considering the majority vote in ensemble defenses is wasteful as it does not effectively utilize available infor mation in the logits layers of the base models. Instead, we propose Run-Off Elec tion (ROE), a novel aggregation method based on a two-round election across the base models: In the first round, models vote for their preferred class and then a second, Run-Off election is held between the top two classes in the first roun d. Based on this approach, we propose DPA+ROE and FA+ROE defense methods based o n Deep Partition Aggregation (DPA) and Finite Aggregation (FA) approaches from p rior work. We evaluate our methods on MNIST, CIFAR-10, and GTSRB and obtain impr ovements in certified accuracy by up to $3%$-$4%$. Also, by applying ROE on a bo osted version of DPA, we gain improvements around $12%$-$27%$ comparing to the c urrent state-of-the-art, establishing a new state-of-the-art in (pointwise) cert ified robustness against data poisoning. In many cases, our approach outperforms the state-of-the-art, even when using 32 times less computational power.
****

Learning Control-Oriented Dynamical Structure from Data
Spencer M. Richards, Jean-Jacques Slotine, Navid Azizan, Marco Pavone
Even for known nonlinear dynamical systems, feedback controller synthesis is a d ifficult problem that often requires leveraging the particular structure of the dynamics to induce a stable closed-loop system. For general nonlinear models, in cluding those fit to data, there may not be enough known structure to reliably s ynthesize a stabilizing feedback controller. In this paper, we discuss a state-d ependent nonlinear tracking controller formulation based on a state-dependent Ri ccati equation for general nonlinear control-affine systems. This formulation de pends on a nonlinear factorization of the system of vector fields defining the c ontrol-affine dynamics, which always exists under mild smoothness assumptions. W e propose a method for learning this factorization from a finite set of data. On a variety of simulated nonlinear dynamical systems, we empirically demonstrate the efficacy of learned versions of this controller in stable trajectory trackin g. Alongside our learning method, we evaluate recent ideas in jointly learning a controller and stabilizability certificate for known dynamical systems; we show experimentally that such methods can be frail in comparison.
****

The Edge of Orthogonality: A Simple View of What Makes BYOL Tick
Pierre Harvey Richemond, Allison Tam, Yunhao Tang, Florian Strub, Bilal Piot, Fe lix Hill
Self-predictive unsupervised learning methods such as BYOL or SimSIAM have shown impressive results, and counter-intuitively, do not collapse to trivial represe ntations. In this work, we aim at exploring the simplest possible mathematical a rguments towards explaining the underlying mechanisms behind self-predictive uns upervised learning. We start with the observation that those methods crucially r ely on the presence of a predictor network (and stop-gradient). With simple line ar algebra, we show that when using a linear predictor, the optimal predictor is close to an orthogonal projection, and propose a general framework based on ort honormalization that enables to interpret and give intuition on why BYOL works. In addition, this framework demonstrates the crucial role of the exponential mov ing average and stop-gradient operator in BYOL as an efficient orthonormalizatio n mechanism. We use these insights to propose four new closed-form predictor var iants of BYOL to support our analysis. Our closed-form predictors outperform sta ndard linear trainable predictor BYOL at 100 and 300 epochs (top-1 linear accura cy on ImageNet).
****

Multi-Agent Best Arm Identification with Private Communications
Alexandre Rio, Merwan Barlier, Igor Colin, Marta Soare
We address multi-agent best arm identification with privacy guarantees. In this setting, agents collaborate by communicating to find the optimal arm. To avoid l

eaking sensitive data through messages, we consider two notions of privacy withh olding different kinds of information: differential privacy and (ϵ, η)-privacy. For each privacy definition, we propose an algorithm based on a two-level successive elimination scheme. We provide theoretical guarantees for the p rivacy level, accuracy and sample complexity of our algorithms. Experiments on v arious settings support our theoretical findings.
****

A Two-Stage Active Learning Algorithm for k-Nearest Neighbors
Nicholas Rittler, Kamalika Chaudhuri
k-nearest neighbor classification is a popular non-parametric method because o f desirable properties like automatic adaption to distributional scale changes. Unfortunately, it has thus far proved difficult to design active learning strate gies for the training of local voting-based classifiers that naturally retain th ese desirable properties, and hence active learning strategies for k-nearest n eighbor classification have been conspicuously missing from the literature. In t his work, we introduce a simple and intuitive active learning algorithm for the training of k-nearest neighbor classifiers, the first in the literature which retains the concept of the k-nearest neighbor vote at prediction time. We prov ide consistency guarantees for a modified k-nearest neighbors classifier train ed on samples acquired via our scheme, and show that when the conditional probab ility function $\mathbb{P}(Y=y|X=x)$ is sufficiently smooth and the Tsybakov noi se condition holds, our actively trained classifiers converge to the Bayes optim al classifier at a faster asymptotic rate than passively trained k-nearest nei ghbor classifiers.
****

Lowering the Pre-training Tax for Gradient-based Subset Training: A Lightweight Distributed Pre-Training Toolkit
Yeonju Ro, Zhangyang Wang, Vijay Chidambaram, Aditya Akella
Training data and model sizes are increasing exponentially. One way to reduce tr aining time and resources is to train with a carefully selected subset of the fu ll dataset. Prior work uses the gradient signals obtained during a warm-up or “p re-training" phase over the full dataset, for determining the core subset; if th e pre-training phase is too small, the gradients obtained are chaotic and unreli able. As a result, the pre-training phase itself incurs significant time/resourc e overhead, and prior work has not gone beyond hyperparameter search to reduce p re-training time. Our work explicitly aims to reduce this $\textbf{pre-training tax}$ in gradient-based subset training. We develop a principled, scalable appro ach for pre-training in a distributed setup. Our approach is $\textit{lightweigh t}$ and $\textit{minimizes communication}$ between distributed worker nodes. It is the first to utilize the concept of model-soup based distributed training $\t extit{at initialization}$. The key idea is to minimally train an ensemble of mod els on small, disjointed subsets of the data; we further employ data-driven spar sity and data augmentation for local worker training to boost ensemble diversity . The centralized model, obtained at the end of pre-training by merging the per-worker models, is found to offer stabilized gradient signals to select subsets, on which the main model is further trained. We have validated the effectiveness of our method through extensive experiments on CIFAR-10/100, and ImageNet, using ResNet and WideResNet models. For example, our approach is shown to achieve $\t extbf{15.4$\times$}$ pre-training speedup and $\textbf{2.8$\times$}$ end-to-end speedup on CIFAR10 and ResNet18 without loss of accuracy. The code is at https:/ /github.com/moonbucks/LiPT.git.
****

The Role of Entropy and Reconstruction in Multi-View Self-Supervised Learning
Borja Rodr■■guez Gálvez, Arno Blaas, Pau Rodriguez, Adam Golinski, Xavier Suau, Jason Ramapuram, Dan Busbridge, Luca Zappella
The mechanisms behind the success of multi-view self-supervised learning (MVSSL) are not yet fully understood. Contrastive MVSSL methods have been studied throu gh the lens of InfoNCE, a lower bound of the Mutual Information (MI). However, t he relation between other MVSSL methods and MI remains unclear. We consider a di fferent lower bound on the MI consisting of an entropy and a reconstruction term

(ER), and analyze the main MVSSL families through its lens. Through this ER bound, we show that clustering-based methods such as DeepCluster and SwAV maximize the MI. We also re-interpret the mechanisms of distillation-based approaches such as BYOL and DINO, showing that they explicitly maximize the reconstruction term and implicitly encourage a stable entropy, and we confirm this empirically. We show that replacing the objectives of common MVSSL methods with this ER bound achieves competitive performance, while making them stable when training with smaller batch sizes or smaller exponential moving average (EMA) coefficients.
****

RLang: A Declarative Language for Describing Partial World Knowledge to Reinforcement Learning Agents
Rafael Rodriguez-Sanchez, Benjamin Adin Spiegel, Jennifer Wang, Roma Patel, Stefanie Tellex, George Konidaris
We introduce RLang, a domain-specific language (DSL) for communicating domain knowledge to an RL agent. Unlike existing RL DSLs that ground to $\textit{single}$ elements of a decision-making formalism (e.g., the reward function or policy), RLang can specify information about every element of a Markov decision process. We define precise syntax and grounding semantics for RLang, and provide a parser that grounds RLang programs to an algorithm-agnostic $\textit{partial}$ world model and policy that can be exploited by an RL agent. We provide a series of example RLang programs demonstrating how different RL methods can exploit the resulting knowledge, encompassing model-free and model-based tabular algorithms, policy gradient and value-based methods, hierarchical approaches, and deep methods.
****

Improving Fair Training under Correlation Shifts
Yuji Roh, Kangwook Lee, Steven Euijong Whang, Changho Suh
Model fairness is an essential element for Trustworthy AI. While many techniques for model fairness have been proposed, most of them assume that the training and deployment data distributions are identical, which is often not true in practice. In particular, when the bias between labels and sensitive groups changes, the fairness of the trained model is directly influenced and can worsen. We make two contributions for solving this problem. First, we analytically show that existing in-processing fair algorithms have fundamental limits in accuracy and group fairness. We utilize the notion of correlation shifts between labels and groups, which can explicitly capture the change of the above bias. Second, we propose a novel pre-processing step that samples the input data to reduce correlation shifts and thus enables the in-processing approaches to overcome their limitations. We formulate an optimization problem for adjusting the data ratio among labels and sensitive groups to reflect the shifted correlation. A key benefit of our approach lies in decoupling the roles of pre- and in-processing approaches: correlation adjustment via pre-processing and unfairness mitigation on the processed data via in-processing. Experiments show that our framework effectively improves existing in-processing fair algorithms w.r.t. accuracy and fairness, both on synthetic and real datasets.
****

The Statistical Benefits of Quantile Temporal-Difference Learning for Value Estimation
Mark Rowland, Yunhao Tang, Clare Lyle, Remi Munos, Marc G Bellemare, Will Dabney
We study the problem of temporal-difference-based policy evaluation in reinforcement learning. In particular, we analyse the use of a distributional reinforcement learning algorithm, quantile temporal-difference learning (QTD), for this task. We reach the surprising conclusion that even if a practitioner has no interest in the return distribution beyond the mean, QTD (which learns predictions about the full distribution of returns) may offer performance superior to approaches such as classical TD learning, which predict only the mean return, even in the tabular setting.
****

Robust Satisficing MDPs
Haolin Ruan, Siyu Zhou, Zhi Chen, Chin Pang Ho
Despite being a fundamental building block for reinforcement learning, Markov de

cision processes (MDPs) often suffer from ambiguity in model parameters. Robust MDPs are proposed to overcome this challenge by optimizing the worst-case performance under ambiguity. While robust MDPs can provide reliable policies with limited data, their worst-case performances are often overly conservative, and so they do not offer practical insights into the actual performance of these reliable policies. This paper proposes robust satisficing MDPs (RSMDPs), where the expected returns of feasible policies are softly-constrained to achieve a user-specified target under ambiguity. We derive a tractable reformulation for RSMDPs and develop a first-order method for solving large instances. Experimental results demonstrate that RSMDPs can prescribe policies to achieve their targets, which are much higher than the optimal worst-case returns computed by robust MDPs. Moreover, the average and percentile performances of our model are competitive among other models. We also demonstrate the scalability of the proposed algorithm compared with a state-of-the-art commercial solver.

****

Infinite Action Contextual Bandits with Reusable Data Exhaust

Mark Rucker, Yinglun Zhu, Paul Mineiro

For infinite action contextual bandits, smoothed regret and reduction to regression results in state-of-the-art online performance with computational cost independent of the action set: unfortunately, the resulting data exhaust does not have well-defined importance-weights. This frustrates the execution of downstream data science processes such as offline model selection. In this paper we describe an online algorithm with an equivalent smoothed regret guarantee, but which generates well-defined importance weights: in exchange, the online computational cost increases, but only to order smoothness (i.e., still independent of the action set). This removes a key obstacle to adoption of smoothed regret in production scenarios.

****

Function-Space Regularization in Neural Networks: A Probabilistic Perspective

Tim G. J. Rudner, Sanyam Kapoor, Shikai Qiu, Andrew Gordon Wilson

Parameter-space regularization in neural network optimization is a fundamental tool for improving generalization. However, standard parameter-space regularization methods make it challenging to encode explicit preferences about desired predictive functions into neural network training. In this work, we approach regularization in neural networks from a probabilistic perspective and show that by viewing parameter-space regularization as specifying an empirical prior distribution over the model parameters, we can derive a probabilistically well-motivated regularization technique that allows explicitly encoding information about desired predictive functions into neural network training. This method—which we refer to as function-space empirical Bayes (FS-EB)—includes both parameter- and function-space regularization, is mathematically simple, easy to implement, and incurs only minimal computational overhead compared to standard regularization techniques. We evaluate the utility of this regularization technique empirically and demonstrate that the proposed method leads to near-perfect semantic shift detection, highly-calibrated predictive uncertainty estimates, successful task adaption from pre-trained models, and improved generalization under covariate shift.

****

A New PHO-rmula for Improved Performance of Semi-Structured Networks

David Rügamer

Recent advances to combine structured regression models and deep neural networks for better interpretability, more expressiveness, and statistically valid uncertainty quantification demonstrate the versatility of semi-structured neural networks (SSNs). We show that techniques to properly identify the contributions of the different model components in SSNs, however, lead to suboptimal network estimation, slower convergence, and degenerated or erroneous predictions. In order to solve these problems while preserving favorable model properties, we propose a non-invasive post-hoc orthogonalization (PHO) that guarantees identifiability of model components and provides better estimation and prediction quality. Our theoretical findings are supported by numerical experiments, a benchmark comparison as well as a real-world application to COVID-19 infections.

****
Geometric Clifford Algebra Networks
David Ruhe, Jayesh K Gupta, Steven De Keninck, Max Welling, Johannes Brandstetter
We propose Geometric Clifford Algebra Networks (GCANs) for modeling dynamical systems. GCANs are based on symmetry group transformations using geometric (Clifford) algebras. We first review the quintessence of modern (plane-based) geometric algebra, which builds on isometries encoded as elements of the $\mathrm{Pin}(p,q,r)$ group. We then propose the concept of group action layers, which linearly combine object transformations using pre-specified group actions. Together with a new activation and normalization scheme, these layers serve as adjustable geometric templates that can be refined via gradient descent. Theoretical advantages are strongly reflected in the modeling of three-dimensional rigid body transformations as well as large-scale fluid dynamics simulations, showing significantly improved performance over traditional methods.
****
Constrained Monotonic Neural Networks
Davor Runje, Sharath M Shankaranarayana
Wider adoption of neural networks in many critical domains such as finance and healthcare is being hindered by the need to explain their predictions and to impose additional constraints on them. Monotonicity constraint is one of the most requested properties in real-world scenarios and is the focus of this paper. One of the oldest ways to construct a monotonic fully connected neural network is to constrain signs on its weights. Unfortunately, this construction does not work with popular non-saturated activation functions as it can only approximate convex functions. We show this shortcoming can be fixed by constructing two additional activation functions from a typical unsaturated monotonic activation function and employing each of them on the part of neurons. Our experiments show this approach of building monotonic neural networks has better accuracy when compared to other state-of-the-art methods, while being the simplest one in the sense of having the least number of parameters, and not requiring any modifications to the learning procedure or post-learning steps. Finally, we prove it can approximate any continuous monotone function on a compact subset of $\mathbb{R}^n$.
****
Differential Privacy, Linguistic Fairness, and Training Data Influence: Impossibility and Possibility Theorems for Multilingual Language Models
Phillip Rust, Anders Søgaard
Language models such as mBERT, XLM-R, and BLOOM aim to achieve multilingual generalization or compression to facilitate transfer to a large number of (potentially unseen) languages. However, these models should ideally also be private, linguistically fair, and transparent, by relating their predictions to training data. Can these requirements be simultaneously satisfied? We show that multilingual compression and linguistic fairness are compatible with differential privacy, but that differential privacy is at odds with training data influence sparsity, an objective for transparency. We further present a series of experiments on two common NLP tasks and evaluate multilingual compression and training data influence sparsity under different privacy guarantees, exploring these trade-offs in more detail. Our results suggest that we need to develop ways to jointly optimize for these objectives in order to find practical trade-offs.
****
Intrinsic Sliced Wasserstein Distances for Comparing Collections of Probability Distributions on Manifolds and Graphs
Raif M. Rustamov, Subhabrata Majumdar
Collections of probability distributions arise in a variety of applications ranging from user activity pattern analysis to brain connectomics. In practice these distributions can be defined over diverse domain types including finite intervals, circles, cylinders, spheres, other manifolds, and graphs. This paper introduces an approach for detecting differences between two collections of distributions over such general domains. To this end, we propose the intrinsic slicing construction that yields a novel class of Wasserstein distances on manifolds and gra

phs. These distances are Hilbert embeddable, allowing us to reduce the distribution collection comparison problem to a more familiar mean testing problem in a Hilbert space. We provide two testing procedures one based on resampling and another on combining p-values from coordinate-wise tests. Our experiments in various synthetic and real data settings show that the resulting tests are powerful and the p-values are well-calibrated.
****

SWARM Parallelism: Training Large Models Can Be Surprisingly Communication-Efficient
Max Ryabinin, Tim Dettmers, Michael Diskin, Alexander Borzunov
Many deep learning applications benefit from using large models with billions of parameters. Training these models is notoriously expensive due to the need for specialized HPC clusters. In this work, we consider alternative setups for training large models: using cheap “preemptible” instances or pooling existing resources from multiple regions. We analyze the performance of existing model-parallel algorithms in these conditions and find configurations where training larger models becomes less communication-intensive. Based on these findings, we propose SWARM Parallelism (Stochastically Wired Adaptively Rebalanced Model Parallelism), a model-parallel training algorithm designed for poorly connected, heterogeneous and unreliable devices. SWARM creates temporary randomized pipelines between nodes that are rebalanced in case of failure. We empirically validate our findings and compare SWARM Parallelism with existing large-scale training approaches. Finally, we combine our insights with compression strategies to train a large Transformer language model with 1B shared parameters ($\approx$13B before sharing) on preemptible T4 GPUs with less than 200 Mb/s network.
****

Hiera: A Hierarchical Vision Transformer without the Bells-and-Whistles
Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, Jitendra Malik, Yanghao Li, Christoph Feichtenhofer
Modern hierarchical vision transformers have added several vision-specific components in the pursuit of supervised classification performance. While these components lead to effective accuracies and attractive FLOP counts, the added complexity actually makes these transformers slower than their vanilla ViT counterparts. In this paper, we argue that this additional bulk is unnecessary. By pretraining with a strong visual pretext task (MAE), we can strip out all the bells-and-whistles from a state-of-the-art multi-stage vision transformer without losing accuracy. In the process, we create Hiera, an extremely simple hierarchical vision transformer that is more accurate than previous models while being significantly faster both at inference and during training. We evaluate Hiera on a variety of tasks for image and video recognition. Our code and models are available at https://github.com/facebookresearch/hiera.
****

End-to-End Learning for Stochastic Optimization: A Bayesian Perspective
Yves Rychener, Daniel Kuhn, Tobias Sutter
We develop a principled approach to end-to-end learning in stochastic optimization. First, we show that the standard end-to-end learning algorithm admits a Bayesian interpretation and trains a posterior Bayes action map. Building on the insights of this analysis, we then propose new end-to-end learning algorithms for training decision maps that output solutions of empirical risk minimization and distributionally robust optimization problems, two dominant modeling paradigms in optimization under uncertainty. Numerical results for a synthetic newsvendor problem illustrate the key differences between alternative training schemes. We also investigate an economic dispatch problem based on real data to showcase the impact of the neural network architecture of the decision maps on their test performance.
****

Sequential Monte Carlo Learning for Time Series Structure Discovery
Feras Saad, Brian Patton, Matthew Douglas Hoffman, Rif A. Saurous, Vikash Mansinghka

This paper presents a new approach to automatically discovering accurate models of complex time series data. Working within a Bayesian nonparametric prior over a symbolic space of Gaussian process time series models, we present a novel structure learning algorithm that integrates sequential Monte Carlo (SMC) and involutive MCMC for highly effective posterior inference. Our method can be used both in "online" settings, where new data is incorporated sequentially in time, and in "offline" settings, by using nested subsets of historical data to anneal the posterior. Empirical measurements on real-world time series show that our method can deliver 10x–100x runtime speedups over previous MCMC and greedy-search structure learning algorithms targeting the same model family. We use our method to perform the first large-scale evaluation of Gaussian process time series structure learning on a prominent benchmark of 1,428 econometric datasets. The results show that our method discovers sensible models that deliver more accurate point forecasts and interval forecasts over multiple horizons as compared to widely used statistical and neural baselines that struggle on this challenging data.

****

Active Ranking of Experts Based on their Performances in Many Tasks

El Mehdi Saad, Nicolas Verzelen, Alexandra Carpentier

We consider the problem of ranking n experts based on their performances on d tasks. We make a monotonicity assumption stating that for each pair of experts, one outperforms the other on all tasks. We consider the sequential setting where in each round the learner has access to noisy evaluations of actively chosen pair of expert-task, given the information available up to the actual round. Given a confidence parameter $\delta \in (0, 1)$, we provide strategies allowing to recover the correct ranking of experts and develop a bound on the total number of queries made by our algorithm that hold with probability at least $1-\delta$. We show that our strategy is adaptive to the complexity of the problem (our bounds are instance dependent), and develop matching lower bounds up to a ploy-logarithmic factor. Finally, we adapt our strategy to the relaxed problem of best expert identification and provide numerical simulation consistent with our theoretical results

****

Sample Complexity Bounds for Learning High-dimensional Simplices in Noisy Regimes

Seyed Amir Hossein Saberi, Amir Najafi, Abolfazl Motahari, Babak Khalaj

In this paper, we propose sample complexity bounds for learning a simplex from noisy samples. A dataset of size n is given which includes i.i.d. samples drawn from a uniform distribution over an unknown arbitrary simplex in $\mathbb{R}^K$, where samples are assumed to be corrupted by a multi-variate additive Gaussian noise of an arbitrary magnitude. We prove the existence of an algorithm that with high probability outputs a simplex having a ℓ_2 distance of at most ε from the true simplex (for any $\varepsilon>0$). Also, we theoretically show that in order to achieve this bound, it is sufficient to have $n\ge\tilde{\Omega}\left(K^2/\varepsilon^2\right)e^{\Omega\left(K/\mathrm{SNR}^2\right)}$ samples, where SNR stands for the signal-to-noise ratio and is defined as the ratio of the maximum component-wise standard deviation of the simplex (signal) to that of the noise vector. This result solves an important open problem in this area of research, and shows as long as $\mathrm{SNR}\ge\Omega\left(\sqrt{K}\right)$ the sample complexity of the noisy regime has the same order to that of the noiseless case. Our proofs are a combination of the so-called sample compression technique in (Ashtiani et al., 2018), mathematical tools from high-dimensional geometry, and Fourier analysis. In particular, we have proposed a general Fourier-based technique for recovery of a more general class of distribution families from additive Gaussian noise, which can be further used in a variety of other related problems.

****

Global Selection of Contrastive Batches via Optimization on Sample Permutations

Vin Sachidananda, Ziyi Yang, Chenguang Zhu

Contrastive Learning has recently achieved state-of-the-art performance in a wide range of unimodal and multimodal tasks. Many contrastive learning approaches u

se mined hard negatives to make batches more informative during training but the
se approaches are inefficient as they increase epoch length proportional to the
number of mined negatives and require frequent updates of nearest neighbor indic
es or mining from recent batches. In this work, we provide an alternative to har
d negative mining, Global Contrastive Batch Sampling (GCBS), an efficient approx
imation to the batch assignment problem that upper bounds the gap between the gl
obal and training losses, $\mathcal{L}^{Global} - \mathcal{L}^{Train}$, in contr
astive learning settings. Through experimentation we find GCBS improves state-of
-the-art performance in sentence embedding and code-search tasks. Additionally,
GCBS is easy to implement as it requires only a few additional lines of code, do
es not maintain external data structures such as nearest neighbor indices, is mo
re computationally efficient than the most minimal hard negative mining approach
es, and makes no changes to the model being trained. Code is available at https:
//github.com/vinayak1/GCBS.
****

High-Probability Bounds for Stochastic Optimization and Variational Inequalities
: the Case of Unbounded Variance

Abdurakhmon Sadiev, Marina Danilova, Eduard Gorbunov, Samuel Horváth, Gauthier G
idel, Pavel Dvurechensky, Alexander Gasnikov, Peter Richtárik

During the recent years the interest of optimization and machine learning commun
ities in high-probability convergence of stochastic optimization methods has bee
n growing. One of the main reasons for this is that high-probability complexity
bounds are more accurate and less studied than in-expectation ones. However, SOT
A high-probability non-asymptotic convergence results are derived under strong a
ssumptions such as boundedness of the gradient noise variance or of the objectiv
e’s gradient itself. In this paper, we propose several algorithms with high-prob
ability convergence results under less restrictive assumptions. In particular, w
e derive new high-probability convergence results under the assumption that the
gradient/operator noise has bounded central α-th moment for $\alpha \in (
1,2]$ in the following setups: (i) smooth non-convex / Polyak-Lojasiewicz / conv
ex / strongly convex / quasi-strongly convex minimization problems, (ii) Lipschi
tz / star-cocoercive and monotone / quasi-strongly monotone variational inequali
ties. These results justify the usage of the considered methods for solving prob
lems that do not fit standard functional classes studied in stochastic optimizat
ion.
****

End-to-end Differentiable Clustering with Associative Memories

Bishwajit Saha, Dmitry Krotov, Mohammed J Zaki, Parikshit Ram

Clustering is a widely used unsupervised learning technique involving an intensi
ve discrete optimization problem. Associative Memory models or AMs are different
iable neural networks defining a recursive dynamical system, which have been int
egrated with various deep learning architectures. We uncover a novel connection
between the AM dynamics and the inherent discrete assignment necessary in cluste
ring to propose a novel unconstrained continuous relaxation of the discrete clus
tering problem, enabling end-to-end differentiable clustering with AM, dubbed Cl
AM. Leveraging the pattern completion ability of AMs, we further develop a novel
 self-supervised clustering loss. Our evaluations on varied datasets demonstrate
 that ClAM benefits from the self-supervision, and significantly improves upon b
oth the traditional Lloyd’s k-means algorithm, and more recent continuous cluste
ring relaxations (by upto 60% in terms of the Silhouette Coefficient).
****

Learning to Suggest Breaks: Sustainable Optimization of Long-Term User Engagemen
t

Eden Saig, Nir Rosenfeld

Optimizing user engagement is a key goal for modern recommendation systems, but
blindly pushing users towards increased consumption risks burn-out, churn, or ev
en addictive habits. To promote digital well-being, most platforms now offer a s
ervice that periodically prompts users to take breaks. These, however, must be s
et up manually, and so may be suboptimal for both users and the system. In this
paper, we study the role of breaks in recommendation, and propose a framework fo

r learning optimal breaking policies that promote and sustain long-term engageme
nt. Based on the notion that recommendation dynamics are susceptible to both pos
itive and negative feedback, we cast recommendation as a Lotka-Volterra dynamica
l system, where breaking reduces to a problem of optimal control. We then give a
n efficient learning algorithm, provide theoretical guarantees, and empirically
demonstrate the utility of our approach on semi-synthetic data.
****

Multi-class Graph Clustering via Approximated Effective p-Resistance
Shota Saito, Mark Herbster
This paper develops an approximation to the (effective) p-resistance and appli
es it to multi-class clustering. Spectral methods based on the graph Laplacian a
nd its generalization to the graph p-Laplacian have been a backbone of non-euc
lidean clustering techniques. The advantage of the p-Laplacian is that the par
ameter p induces a controllable bias on cluster structure. The drawback of p
-Laplacian eigenvector based methods is that the third and higher eigenvectors a
re difficult to compute. Thus, instead, we are motivated to use the p-resistan
ce induced by the p-Laplacian for clustering. For p-resistance, small p bi
ases towards clusters with high internal connectivity while large p biases tow
ards clusters of small “extent,” that is a preference for smaller shortest-path
distances between vertices in the cluster. However, the p-resistance is expens
ive to compute. We overcome this by developing an approximation to the p-resis
tance. We prove upper and lower bounds on this approximation and observe that it
 is exact when the graph is a tree. We also provide theoretical justification fo
r the use of p-resistance for clustering. Finally, we provide experiments comp
aring our approximated p-resistance clustering to other p-Laplacian based me
thods.
****

Off-Policy Evaluation for Large Action Spaces via Conjunct Effect Modeling
Yuta Saito, Qingyang Ren, Thorsten Joachims
We study off-policy evaluation (OPE) of contextual bandit policies for large dis
crete action spaces where conventional importance-weighting approaches suffer fr
om excessive variance. To circumvent this variance issue, we propose a new estim
ator, called OffCEM, that is based on the conjunct effect model (CEM), a novel d
ecomposition of the causal effect into a cluster effect and a residual effect. O
ffCEM applies importance weighting only to action clusters and addresses the res
idual causal effect through model-based reward estimation. We show that the prop
osed estimator is unbiased under a new assumption, called local correctness, whi
ch only requires that the residual-effect model preserves the relative expected
reward differences of the actions within each cluster. To best leverage the CEM
and local correctness, we also propose a new two-step procedure for performing m
odel-based estimation that minimizes bias in the first step and variance in the
second step. We find that the resulting OffCEM estimator substantially improves
bias and variance compared to a range of conventional estimators. Experiments de
monstrate that OffCEM provides substantial improvements in OPE especially in the
 presence of many actions.
****

Rethinking Warm-Starts with Predictions: Learning Predictions Close to Sets of O
ptimal Solutions for Faster L-/$\text{L}^\natural$-Convex Function Mini
mization
Shinsaku Sakaue, Taihei Oki
An emerging line of work has shown that machine-learned predictions are useful t
o warm-start algorithms for discrete optimization problems, such as bipartite ma
tching. Previous studies have shown time complexity bounds proportional to some
distance between a prediction and an optimal solution, which we can approximatel
y minimize by learning predictions from past optimal solutions. However, such gu
arantees may not be meaningful when multiple optimal solutions exist. Indeed, th
e dual problem of bipartite matching and, more generally, L-/$\text{L}^
\\natural$-convex function minimization have arbitrarily many optimal solutions,
 making such prediction-dependent bounds arbitrarily large. To resolve this theo
retically critical issue, we present a new warm-start-with-prediction framework

for L-/$\text{L}^\\natural$-convex function minimization. Our framework offers time complexity bounds proportional to the distance between a prediction and the set of all optimal solutions. The main technical difficulty lies in learning predictions that are provably close to sets of all optimal solutions, for which we present an online-gradient-descent-based method. We thus give the first polynomial-time learnability of predictions that can provably warm-start algorithms regardless of multiple optimal solutions.

****

PAC-Bayesian Offline Contextual Bandits With Guarantees

Otmane Sakhi, Pierre Alquier, Nicolas Chopin

This paper introduces a new principled approach for off-policy learning in contextual bandits. Unlike previous work, our approach does not derive learning principles from intractable or loose bounds. We analyse the problem through the PAC-Bayesian lens, interpreting policies as mixtures of decision rules. This allows us to propose novel generalization bounds and provide tractable algorithms to optimize them. We prove that the derived bounds are tighter than their competitors, and can be optimized directly to confidently improve upon the logging policy offline. Our approach learns policies with guarantees, uses all available data and does not require tuning additional hyperparameters on held-out sets. We demonstrate through extensive experiments the effectiveness of our approach in providing performance guarantees in practical scenarios.

****

Provably and Practically Efficient Neural Contextual Bandits

Sudeep Salgia

We consider the neural contextual bandit problem. In contrast to the existing work which primarily focuses on ReLU neural nets, we consider a general set of smooth activation functions. Under this more general setting, (i) we derive non-asymptotic error bounds on the difference between an overparameterized neural net and its corresponding neural tangent kernel, (ii) we propose an algorithm with a provable sublinear regret bound that is also efficient in the finite regime as demonstrated by empirical studies. The non-asymptotic error bounds may be of broader interests as a tool to establish the relation between the smoothness of the activation functions in neural contextual bandits and the smoothness of the kernels in kernel bandits.

****

Distributed Linear Bandits under Communication Constraints

Sudeep Salgia, Qing Zhao

We consider distributed linear bandits where M agents learn collaboratively to minimize the overall cumulative regret incurred by all agents. Information exchange is facilitated by a central server, and both the uplink and downlink communications are carried over channels with fixed capacity, which limits the amount of information that can be transmitted in each use of the channels. We investigate the regret-communication trade-off by (i) establishing information-theoretic lower bounds on the required communications (in terms of bits) for achieving a sublinear regret order; (ii) developing an efficient algorithm that achieves the minimum sublinear regret order offered by centralized learning using the minimum order of communications dictated by the information-theoretic lower bounds. For sparse linear bandits, we show a variant of the proposed algorithm offers better regret-communication trade-off by leveraging the sparsity of the problem.

****

Optimizing Hyperparameters with Conformal Quantile Regression

David Salinas, Jacek Golebiowski, Aaron Klein, Matthias Seeger, Cedric Archambeau

Many state-of-the-art hyperparameter optimization (HPO) algorithms rely on model-based optimizers that learn surrogate models of the target function to guide the search. Gaussian processes are the de facto surrogate model due to their ability to capture uncertainty. However, they make strong assumptions about the observation noise, which might not be warranted in practice. In this work, we propose to leverage conformalized quantile regression which makes minimal assumptions about the observation noise and, as a result, models the target function in a mor

e realistic and robust fashion which translates to quicker HPO convergence on em pirical benchmarks. To apply our method in a multi-fidelity setting, we propose a simple, yet effective, technique that aggregates observed results across diffe rent resource levels and outperforms conventional methods across many empirical tasks.

****

Raising the Cost of Malicious AI-Powered Image Editing

Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, Aleksander Madry

We present an approach to mitigating the risks of malicious image editing posed by large diffusion models. The key idea is to immunize images so as to make them resistant to manipulation by these models. This immunization relies on injectio n of imperceptible adversarial perturbations designed to disrupt the operation o f the targeted diffusion models, forcing them to generate unrealistic images. We provide two methods for crafting such perturbations, and then demonstrate their efficacy. Finally, we discuss a policy component necessary to make our approach fully effective and practical—one that involves the organizations developing di ffusion models, rather than individual users, to implement (and support) the imm unization process.

****

Fast, Differentiable and Sparse Top-k: a Convex Analysis Perspective

Michael Eli Sander, Joan Puigcerver, Josip Djolonga, Gabriel Peyré, Mathieu Blon del

The top-k operator returns a k-sparse vector, where the non-zero values corr espond to the k largest values of the input. Unfortunately, because it is a di scontinuous function, it is difficult to incorporate in neural networks trained end-to-end with backpropagation. Recent works have considered differentiable rel axations, based either on regularization or perturbation techniques. However, to date, no approach is fully differentiable and sparse. In this paper, we propose new differentiable and sparse top-k operators. We view the top-k operator a s a linear program over the permutahedron, the convex hull of permutations. We t hen introduce a p-norm regularization term to smooth out the operator, and sho w that its computation can be reduced to isotonic optimization. Our framework is significantly more general than the existing one and allows for example to expr ess top-k operators that select values in magnitude. On the algorithmic side, in addition to pool adjacent violator (PAV) algorithms, we propose a new GPU/TPU -friendly Dykstra algorithm to solve isotonic optimization problems. We successf ully use our operators to prune weights in neural networks, to fine-tune vision transformers, and as a router in sparse mixture of experts.

****

TAN Without a Burn: Scaling Laws of DP-SGD

Tom Sander, Pierre Stock, Alexandre Sablayrolles

Differentially Private methods for training Deep Neural Networks (DNNs) have pro gressed recently, in particular with the use of massive batches and aggregated d ata augmentations for a large number of training steps. These techniques require much more computing resources than their non-private counterparts, shifting the traditional privacy-accuracy trade-off to a privacy-accuracy-compute trade-off and making hyper-parameter search virtually impossible for realistic scenarios. In this work, we decouple privacy analysis and experimental behavior of noisy tr aining to explore the trade-off with minimal computational requirements. We firs t use the tools of Renyi Differential Privacy (RDP) to highlight that the privac y budget, when not overcharged, only depends on the total amount of noise (TAN) injected throughout training. We then derive scaling laws for training models wi th DP-SGD to optimize hyper-parameters with more than a $100\times$ reduction in computational budget. We apply the proposed method on CIFAR-10 and ImageNet and , in particular, strongly improve the state-of-the-art on ImageNet with a $+9$ p oints gain in top-1 accuracy for a privacy budget $\varepsilon=8$.

****

Discrete Continuous Optimization Framework for Simultaneous Clustering and Train ing in Mixture Models

Parth Vipul Sangani, Arjun Shashank Kashettiwar, Pritish Chakraborty, Bhuvan Red

dy Gangula, Durga S, Ganesh Ramakrishnan, Rishabh K Iyer, Abir De
We study a new framework of learning mixture models via automatic clustering called PRESTO, wherein we optimize a joint objective function on the model parameters and the partitioning, with each model tailored to perform well on its specific cluster. In contrast to prior work, we do not assume any generative model for the data. We convert our training problem to a joint parameter estimation cum a subset selection problem, subject to a matroid span constraint. This allows us to reduce our problem into a constrained set function minimization problem, where the underlying objective is monotone and approximately submodular. We then propose a new joint discrete-continuous optimization algorithm that achieves a bounded approximation guarantee for our problem. We show that PRESTO outperforms several alternative methods. Finally, we study PRESTO in the context of resource-efficient deep learning, where we train smaller resource-constrained models on each partition and show that it outperforms existing data partitioning and model pruning/knowledge distillation approaches, which in contrast to PRESTO, require large initial (teacher) models.
****

Whose Opinions Do Language Models Reflect?
Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, Tatsunori Hashimoto
Language models (LMs) are increasingly being used in open-ended contexts, where the opinions they reflect in response to subjective queries can have a profound impact, both on user satisfaction, and shaping the views of society at large. We put forth a quantitative framework to investigate the opinions reflected by LMs – by leveraging high-quality public opinion polls. Using this framework, we create OpinionQA, a dataset for evaluating the alignment of LM opinions with those of 60 US demographic groups over topics ranging from abortion to automation. Across topics, we find substantial misalignment between the views reflected by current LMs and those of US demographic groups: on par with the Democrat-Republican divide on climate change. Notably, this misalignment persists even after explicitly steering the LMs towards particular groups. Our analysis not only confirms prior observations about the left-leaning tendencies of some human feedback-tuned LMs, but also surfaces groups whose opinions are poorly reflected by current LMs (e.g., 65+ and widowed individuals).
****

Streaming Active Learning with Deep Neural Networks
Akanksha Saran, Safoora Yousefi, Akshay Krishnamurthy, John Langford, Jordan T. Ash
Active learning is perhaps most naturally posed as an online learning problem. However, prior active learning approaches with deep neural networks assume offline access to the entire dataset ahead of time. This paper proposes VeSSAL, a new algorithm for batch active learning with deep neural networks in streaming settings, which samples groups of points to query for labels at the moment they are encountered. Our approach trades off between uncertainty and diversity of queried samples to match a desired query rate without requiring any hand-tuned hyperparameters. Altogether, we expand the applicability of deep neural networks to realistic active learning scenarios, such as applications relevant to HCI and large, fractured datasets.
****

Random Teachers are Good Teachers
Felix Sarnthein, Gregor Bachmann, Sotiris Anagnostidis, Thomas Hofmann
In this work, we investigate the implicit regularization induced by teacher-student learning dynamics in self-distillation. To isolate its effect, we describe a simple experiment where we consider teachers at random initialization instead of trained teachers. Surprisingly, when distilling a student into such a random teacher, we observe that the resulting model and its representations already possess very interesting characteristics; (1) we observe a strong improvement of the distilled student over its teacher in terms of probing accuracy. (2) The learned representations are data-dependent and transferable between different tasks but deteriorate strongly if trained on random inputs. (3) The student checkpoint c

ontains sparse subnetworks, so-called lottery tickets, and lies on the border of linear basins in the supervised loss landscape. These observations have interesting consequences for several important areas in machine learning: (1) Self-distillation can work solely based on the implicit regularization present in the gradient dynamics without relying on any dark knowledge, (2) self-supervised learning can learn features even in the absence of data augmentation and (3) training dynamics during the early phase of supervised training do not necessarily require label information. Finally, we shed light on an intriguing local property of the loss landscape: the process of feature learning is strongly amplified if the student is initialized closely to the teacher. These results raise interesting questions about the nature of the landscape that have remained unexplored so far. Code is available at https://github.com/safelix/dinopl.

****

Posterior Sampling for Deep Reinforcement Learning

Remo Sasso, Michelangelo Conserva, Paulo Rauber

Despite remarkable successes, deep reinforcement learning algorithms remain sample inefficient: they require an enormous amount of trial and error to find good policies. Model-based algorithms promise sample efficiency by building an environment model that can be used for planning. Posterior Sampling for Reinforcement Learning is such a model-based algorithm that has attracted significant interest due to its performance in the tabular setting. This paper introduces Posterior Sampling for Deep Reinforcement Learning (PSDRL), the first truly scalable approximation of Posterior Sampling for Reinforcement Learning that retains its model-based essence. PSDRL combines efficient uncertainty quantification over latent state space models with a specially tailored incremental planning algorithm based on value-function approximation. Extensive experiments on the Atari benchmark show that PSDRL significantly outperforms previous state-of-the-art attempts at scaling up posterior sampling while being competitive with a state-of-the-art (model-based) reinforcement learning method, both in sample efficiency and computational efficiency.

****

Graph Neural Networks can Recover the Hidden Features Solely from the Graph Structure

Ryoma Sato

Graph Neural Networks (GNNs) are popular models for graph learning problems. GNNs show strong empirical performance in many practical tasks. However, the theoretical properties have not been completely elucidated. In this paper, we investigate whether GNNs can exploit the graph structure from the perspective of the expressive power of GNNs. In our analysis, we consider graph generation processes that are controlled by hidden (or latent) node features, which contain all information about the graph structure. A typical example of this framework is kNN graphs constructed from the hidden features. In our main results, we show that GNNs can recover the hidden node features from the input graph alone, even when all node features, including the hidden features themselves and any indirect hints, are unavailable. GNNs can further use the recovered node features for downstream tasks. These results show that GNNs can fully exploit the graph structure by themselves, and in effect, GNNs can use both the hidden and explicit node features for downstream tasks. In the experiments, we confirm the validity of our results by showing that GNNs can accurately recover the hidden features using a GNN architecture built based on our theoretical analysis.

****

Existence and Estimation of Critical Batch Size for Training Generative Adversarial Networks with Two Time-Scale Update Rule

Naoki Sato, Hideaki Iiduka

Previous results have shown that a two time-scale update rule (TTUR) using different learning rates, such as different constant rates or different decaying rates, is useful for training generative adversarial networks (GANs) in theory and in practice. Moreover, not only the learning rate but also the batch size is important for training GANs with TTURs and they both affect the number of steps needed for training. This paper studies the relationship between batch size and the

number of steps needed for training GANs with TTURs based on constant learning r
ates. We theoretically show that, for a TTUR with constant learning rates, the n
umber of steps needed to find stationary points of the loss functions of both th
e discriminator and generator decreases as the batch size increases and that the
re exists a critical batch size minimizing the stochastic first-order oracle (SF
O) complexity. Then, we use the Fréchet inception distance (FID) as the performa
nce measure for training and provide numerical results indicating that the numbe
r of steps needed to achieve a low FID score decreases as the batch size increas
es and that the SFO complexity increases once the batch size exceeds the measure
d critical batch size. Moreover, we show that measured critical batch sizes are
close to the sizes estimated from our theoretical results.
****

StyleGAN-T: Unlocking the Power of GANs for Fast Large-Scale Text-to-Image Synth
esis
Axel Sauer, Tero Karras, Samuli Laine, Andreas Geiger, Timo Aila
Text-to-image synthesis has recently seen significant progress thanks to large p
retrained language models, large-scale training data, and the introduction of sc
alable model families such as diffusion and autoregressive models. However, the
best-performing models require iterative evaluation to generate a single sample.
 In contrast, generative adversarial networks (GANs) only need a single forward
pass. They are thus much faster, but they currently remain far behind the state-
of-the-art in large-scale text-to-image synthesis. This paper aims to identify t
he necessary steps to regain competitiveness. Our proposed model, StyleGAN-T, ad
dresses the specific requirements of large-scale text-to-image synthesis, such a
s large capacity, stable training on diverse datasets, strong text alignment, an
d controllable variation vs. text alignment tradeoff. StyleGAN-T significantly i
mproves over previous GANs and outperforms distilled diffusion models - the prev
ious state-of-the-art in fast text-to-image synthesis - in terms of sample quali
ty and speed.
****

Facial Expression Recognition with Adaptive Frame Rate based on Multiple Testing
 Correction
Andrey Savchenko
In this paper, we consider the problem of the high computational complexity of v
ideo-based facial expression recognition. A novel sequential procedure is propos
ed with an adaptive frame rate selection in a short video fragment to speed up d
ecision-making. We automatically adjust the frame rate and process fewer frames
with a low frame rate for more straightforward videos and more frames for comple
x ones. To determine the frame rate at which an inference is sufficiently reliab
le, the Benjamini-Hochberg procedure from multiple comparisons theory is employe
d to control the false discovery rate. The main advantages of our method are an
improvement of the trustworthiness of decision-making by maintaining only one hy
per-parameter (false acceptance rate) and its applicability with arbitrary neura
l network models used as facial feature extractors without the need to re-train
these models. An experimental study on datasets from ABAW and EmotiW challenges
proves the superior performance (1.5-40 times faster) of the proposed approach c
ompared to processing all frames and existing techniques with early exiting and
adaptive frame selection.
****

Off-Policy Average Reward Actor-Critic with Deterministic Policy Search
Naman Saxena, Subhojyoti Khastagir, Shishir Kolathaya, Shalabh Bhatnagar
The average reward criterion is relatively less studied as most existing works i
n the Reinforcement Learning literature consider the discounted reward criterion
. There are few recent works that present on-policy average reward actor-critic
algorithms, but average reward off-policy actor-critic is relatively less explor
ed. In this work, we present both on-policy and off-policy deterministic policy
gradient theorems for the average reward performance criterion. Using these theo
rems, we also present an Average Reward Off-Policy Deep Deterministic Policy Gra
dient (ARO-DDPG) Algorithm. We first show asymptotic convergence analysis using
the ODE-based method. Subsequently, we provide a finite time analysis of the res

ulting stochastic approximation scheme with linear function approximator and obt ain an ϵ-optimal stationary policy with a sample complexity of $\Omega(\epsilon^{-2.5})$. We compare the average reward performance of our proposed ARO -DDPG algorithm and observe better empirical performance compared to state-of-th e-art on-policy average reward actor-critic algorithms over MuJoCo-based environ ments.
****

Gibbsian Polar Slice Sampling
Philip Schär, Michael Habeck, Daniel Rudolf
Polar slice sampling (Roberts & Rosenthal, 2002) is a Markov chain approach for approximate sampling of distributions that is difficult, if not impossible, to i mplement efficiently, but behaves provably well with respect to the dimension. B y updating the directional and radial components of chain iterates separately, w e obtain a family of samplers that mimic polar slice sampling, and yet can be im plemented efficiently. Numerical experiments in a variety of settings indicate t hat our proposed algorithm outperforms the two most closely related approaches, elliptical slice sampling (Murray et al., 2010) and hit-and-run uniform slice sa mpling (MacKay, 2003). We prove the well-definedness and convergence of our meth ods under suitable assumptions on the target distribution.
****

Identifiability and Generalizability in Constrained Inverse Reinforcement Learni ng
Andreas Schlaginhaufen, Maryam Kamgarpour
Two main challenges in Reinforcement Learning (RL) are designing appropriate rew ard functions and ensuring the safety of the learned policy. To address these ch allenges, we present a theoretical framework for Inverse Reinforcement Learning (IRL) in constrained Markov decision processes. From a convex-analytic perspecti ve, we extend prior results on reward identifiability and generalizability to bo th the constrained setting and a more general class of regularizations. In parti cular, we show that identifiability up to potential shaping (Cao et al., 2021) i s a consequence of entropy regularization and may generally no longer hold for o ther regularizations or in the presence of safety constraints. We also show that to ensure generalizability to new transition laws and constraints, the true rew ard must be identified up to a constant. Additionally, we derive a finite sample guarantee for the suboptimality of the learned rewards, and validate our result s in a gridworld environment.
****

Learning Expressive Priors for Generalization and Uncertainty Estimation in Neur al Networks
Dominik Schnaus, Jongseok Lee, Daniel Cremers, Rudolph Triebel
In this work, we propose a novel prior learning method for advancing generalizat ion and uncertainty estimation in deep neural networks. The key idea is to explo it scalable and structured posteriors of neural networks as informative priors w ith generalization guarantees. Our learned priors provide expressive probabilist ic representations at large scale, like Bayesian counterparts of pre-trained mod els on ImageNet, and further produce non-vacuous generalization bounds. We also extend this idea to a continual learning framework, where the favorable properti es of our priors are desirable. Major enablers are our technical contributions: (1) the sums-of-Kronecker-product computations, and (2) the derivations and opti mizations of tractable objectives that lead to improved generalization bounds. E mpirically, we exhaustively show the effectiveness of this method for uncertaint y estimation and generalization.
****

Deterministic equivalent and error universality of deep random features learning
Dominik Schröder, Hugo Cui, Daniil Dmitriev, Bruno Loureiro
This manuscript considers the problem of learning a random Gaussian network func tion using a fully connected network with frozen intermediate layers and trainab le readout layer. This problem can be seen as a natural generalization of the wi dely studied random features model to deeper architectures. First, we prove Gaus sian universality of the test error in a ridge regression setting where the lear

ner and target networks share the same intermediate layers, and provide a sharp asymptotic formula for it. Establishing this result requires proving a deterministic equivalent for traces of the deep random features sample covariance matrices which can be of independent interest. Second, we conjecture the asymptotic Gaussian universality of the test error in the more general setting of arbitrary convex losses and generic learner/target architectures. We provide extensive numerical evidence for this conjecture, which requires the derivation of closed-form expressions for the layer-wise post-activation population covariances. In light of our results, we investigate the interplay between architecture design and implicit regularization.
****

The Acquisition of Physical Knowledge in Generative Neural Networks
Luca M. Schulze Buschoff, Eric Schulz, Marcel Binz
As children grow older, they develop an intuitive understanding of the physical processes around them. Their physical understanding develops in stages, moving along developmental trajectories which have been mapped out extensively in previous empirical research. Here, we investigate how the learning trajectories of deep generative neural networks compare to children's developmental trajectories using physical understanding as a testbed. We outline an approach that allows us to examine two distinct hypotheses of human development – stochastic optimization and complexity increase. We find that while our models are able to accurately predict a number of physical processes, their learning trajectories under both hypotheses do not follow the developmental trajectories of children.
****

Modality-Agnostic Variational Compression of Implicit Neural Representations
Jonathan Richard Schwarz, Jihoon Tack, Yee Whye Teh, Jaeho Lee, Jinwoo Shin
We introduce a modality-agnostic neural compression algorithm based on a functional view of data and parameterised as an Implicit Neural Representation (INR). Bridging the gap between latent coding and sparsity, we obtain compact latent representations non-linearly mapped to a soft gating mechanism. This allows the specialisation of a shared INR network to each data item through subnetwork selection. After obtaining a dataset of such latent representations, we directly optimise the rate/distortion trade-off in a modality-agnostic space using neural compression. Variational Compression of Implicit Neural Representations (VC-INR) shows improved performance given the same representational capacity pre quantisation while also outperforming previous quantisation schemes used for other INR techniques.Our experiments demonstrate strong results over a large set of diverse modalities using the same algorithm without any modality-specific inductive biases. We show results on images, climate data, 3D shapes and scenes as well as audio and video, introducing VC-INR as the first INR-based method to outperform codecs as well-known and diverse as JPEG 2000, MP3 and AVC/HEVC on their respective modalities.
****

Bigger, Better, Faster: Human-level Atari with human-level efficiency
Max Schwarzer, Johan Samir Obando Ceron, Aaron Courville, Marc G Bellemare, Rishabh Agarwal, Pablo Samuel Castro
We introduce a value-based RL agent, which we call BBF, that achieves super-human performance in the Atari 100K benchmark. BBF relies on scaling the neural networks used for value estimation, as well as a number of other design choices that enable this scaling in a sample-efficient manner. We conduct extensive analyses of these design choices and provide insights for future work. We end with a discussion about updating the goalposts for sample-efficient RL research on the ALE. We make our code and data publicly available at https://github.com/google-research/google-research/tree/master/bigger_better_faster.
****

Dissecting the Effects of SGD Noise in Distinct Regimes of Deep Learning
Antonio Sclocchi, Mario Geiger, Matthieu Wyart
Understanding when the noise in stochastic gradient descent (SGD) affects generalization of deep neural networks remains a challenge, complicated by the fact that networks can operate in distinct training regimes. Here we study how the magn

itude of this noise T affects performance as the size of the training set P and the scale of initialization α are varied. For gradient descent, α is a key parameter that controls if the network is lazy' ($\alpha\gg1$) or instead learns features ($\alpha\ll1$). For classification of MNIST and CIFAR10 images, our central results are: *(i)* obtaining phase diagrams for performance in the (α,T) plane. They show that SGD noise can be detrimental or instead useful depending on the training regime. Moreover, although increasing T or decreasing α both allow the net to escape the lazy regime, these changes can have opposite effects on performance. *(ii)* Most importantly, we find that the characteristic temperature T_c where the noise of SGD starts affecting the trained model (and eventually performance) is a power law of P. We relate this finding with the observation that key dynamical quantities, such as the total variation of weights during training, depend on both T and P as power laws. These results indicate that a key effect of SGD noise occurs late in training, by affecting the stopping process whereby all data are fitted. Indeed, we argue that due to SGD noise, nets must develop a strongersignal', i.e. larger informative weights, to fit the data, leading to a longer training time. A stronger signal and a longer training time are also required when the size of the training set P increases. We confirm these views in the perceptron model, where signal and noise can be precisely measured. Interestingly, exponents characterizing the effect of SGD depend on the density of data near the decision boundary, as we explain.

****

A Fast Optimistic Method for Monotone Variational Inequalities

Michael Sedlmayer, Dang-Khoa Nguyen, Radu Ioan Bot

We study monotone variational inequalities that can arise as optimality conditions for constrained convex optimization or convex-concave minimax problems and propose a novel algorithm that uses only one gradient/operator evaluation and one projection onto the constraint set per iteration. The algorithm, which we call fOGDA-VI, achieves a $o(\frac{1}{k})$ rate of convergence in terms of the restricted gap function as well as the natural residual for the last iterate. Moreover, we provide a convergence guarantee for the sequence of iterates to a solution of the variational inequality. These are the best theoretical convergence results for numerical methods for (only) monotone variational inequalities reported in the literature. To empirically validate our algorithm we investigate a two-player matrix game with mixed strategies of the two players. Concluding, we show promising results regarding the application of fOGDA-VI to the training of generative adversarial nets.

****

Double-Weighting for Covariate Shift Adaptation

José I. Segovia-Martín, Santiago Mazuelas, Anqi Liu

Supervised learning is often affected by a covariate shift in which the marginal distributions of instances (covariates x) of training and testing samples $p_\text{tr}(x)$ and $p_\text{te}(x)$ are different but the label conditionals coincide. Existing approaches address such covariate shift by either using the ratio $p_\text{te}(x)/p_\text{tr}(x)$ to weight training samples (reweighted methods) or using the ratio $p_\text{tr}(x)/p_\text{te}(x)$ to weight testing samples (robust methods). However, the performance of such approaches can be poor under support mismatch or when the above ratios take large values. We propose a minimax risk classification (MRC) approach for covariate shift adaptation that avoids such limitations by weighting both training and testing samples. In addition, we develop effective techniques that obtain both sets of weights and generalize the conventional kernel mean matching method. We provide novel generalization bounds for our method that show a significant increase in the effective sample size compared with reweighted methods. The proposed method also achieves enhanced classification performance in both synthetic and empirical experiments.

****

Enhancing Activity Prediction Models in Drug Discovery with the Ability to Understand Human Language

Philipp Seidl, Andreu Vall, Sepp Hochreiter, Günter Klambauer

Activity and property prediction models are the central workhorses in drug discovery and materials sciences, but currently, they have to be trained or fine-tuned for new tasks. Without training or fine-tuning, scientific language models could be used for such low-data tasks through their announced zero- and few-shot capabilities. However, their predictive quality at activity prediction is lacking. In this work, we envision a novel type of activity prediction model that is able to adapt to new prediction tasks at inference time, via understanding textual information describing the task. To this end, we propose a new architecture with separate modules for chemical and natural language inputs, and a contrastive pretraining objective on data from large biochemical databases. In extensive experiments, we show that our method CLAMP yields improved predictive performance on few-shot learning benchmarks and zero-shot problems in drug discovery. We attribute the advances of our method to the modularized architecture and to our pre-training objective.

****

Variational Autoencoding Neural Operators

Jacob H Seidman, Georgios Kissas, George J. Pappas, Paris Perdikaris

Unsupervised learning with functional data is an emerging paradigm of machine learning research with applications to computer vision, climate modeling and physical systems. A natural way of modeling functional data is by learning operators between infinite dimensional spaces, leading to discretization invariant representations that scale independently of the sample grid resolution. Here we present Variational Autoencoding Neural Operators (VANO), a general strategy for making a large class of operator learning architectures act as variational autoencoders. For this purpose, we provide a novel rigorous mathematical formulation of the variational objective in function spaces for training. VANO first maps an input function to a distribution over a latent space using a parametric encoder and then decodes a sample from the latent distribution to reconstruct the input, as in classic variational autoencoders. We test VANO with different model set-ups and architecture choices for a variety of benchmarks. We start from a simple Gaussian random field where we can analytically track what the model learns and progressively transition to more challenging benchmarks including modeling phase separation in Cahn-Hilliard systems and real world satellite data for measuring Earth surface deformation.

****

Neural Markov Jump Processes

Patrick Seifner, Ramses J Sanchez

Markov jump processes are continuous-time stochastic processes with a wide range of applications in both natural and social sciences. Despite their widespread use, inference in these models is highly non-trivial and typically proceeds via either Monte Carlo or expectation-maximization methods. In this work we introduce an alternative, variational inference algorithm for Markov jump processes which relies on neural ordinary differential equations, and is trainable via back-propagation. Our methodology learns neural, continuous-time representations of the observed data, that are used to approximate the initial distribution and time-dependent transition probability rates of the posterior Markov jump process. The time-independent rates of the prior process are in contrast trained akin to generative adversarial networks. We test our approach on synthetic data sampled from ground-truth Markov jump processes, experimental switching ion channel data and molecular dynamics simulations. Source code to reproduce our experiments is available online.

****

Bayesian online change point detection with Hilbert space approximate Student-t process

Jeremy Sellier, Petros Dellaportas

In this paper, we introduce a variant of Bayesian online change point detection with a reducedrank Student-t process (TP) and dependent Student-t noise, as a nonparametric time series model. Our method builds and improves upon the state-of-the-art Gaussian process (GP) change point model benchmark of Saatci et al. (2010). The Student-t process generalizes the concept of a GP and hence yields a mor

e flexible alternative. Additionally, unlike a GP, the predictive variance explicitly depends on the training observations, while the use of an entangled Student-t noise model preserves analytical tractability. Our approach also uses a Hilbert space reduced-rank representation of the TP kernel, derived from an eigenfunction expansion of the Laplace operator (Solin & Sarkka, 2020), to alleviate its computational complexity. Improvements in prediction and training time are demonstrated with real-world data-sets
****
Incentivizing Exploration with Linear Contexts and Combinatorial Actions
Mark Sellke
We advance the study of incentivized bandit exploration, in which arm choices are viewed as recommendations and are required to be Bayesian incentive compatible. Recent work of Sellke-Slivkins (Operations Research 2022) has shown that for the special case of independent arms, after collecting enough initial samples, the popular Thompson sampling algorithm becomes incentive compatible. This was generalized to the combinatorial semibandit in Hu-Ngo-Slivkins-Wu (NeurIPS 2022). We give an analog of this result for linear bandits, where the independence of the prior is replaced by a natural convexity condition. This opens up the possibility of efficient and regret-optimal incentivized exploration in high-dimensional action spaces. In the semibandit model, we also improve the sample complexity for the pre-Thompson sampling phase of initial data collection.
****
Explainability as statistical inference
Hugo Henri Joseph Senetaire, Damien Garreau, Jes Frellsen, Pierre-Alexandre Mattei
A wide variety of model explanation approaches have been proposed in recent years, all guided by very different rationales and heuristics. In this paper, we take a new route and cast interpretability as a statistical inference problem. We propose a general deep probabilistic model designed to produce interpretable predictions. The model’s parameters can be learned via maximum likelihood, and the method can be adapted to any predictor network architecture, and any type of prediction problem. Our model is akin to amortized interpretability methods, where a neural network is used as a selector to allow for fast interpretation at inference time. Several popular interpretability methods are shown to be particular cases of regularized maximum likelihood for our general model. Using our framework, we identify imputation as a common issue of these models. We propose new datasets with ground truth selection which allow for the evaluation of the features importance map and show experimentally that multiple imputation provides more reasonable interpretations.
****
Multi-View Masked World Models for Visual Robotic Manipulation
Younggyo Seo, Junsu Kim, Stephen James, Kimin Lee, Jinwoo Shin, Pieter Abbeel
Visual robotic manipulation research and applications often use multiple cameras, or views, to better perceive the world. How else can we utilize the richness of multi-view data? In this paper, we investigate how to learn good representations with multi-view data and utilize them for visual robotic manipulation. Specifically, we train a multi-view masked autoencoder which reconstructs pixels of randomly masked viewpoints and then learn a world model operating on the representations from the autoencoder. We demonstrate the effectiveness of our method in a range of scenarios, including multi-view control and single-view control with auxiliary cameras for representation learning. We also show that the multi-view masked autoencoder trained with multiple randomized viewpoints enables training a policy with strong viewpoint randomization and transferring the policy to solve real-robot tasks without camera calibration and an adaptation procedure. Video demonstrations are available at: https://sites.google.com/view/mv-mwm.
****
One-Shot Compression of Large Edge-Exchangeable Graphs using Bits-Back Coding
Daniel Severo, James Townsend, Ashish J Khisti, Alireza Makhzani
We present a one-shot method for compressing large labeled graphs called Random Edge Coding. When paired with a parameter-free model based on Pólya’s Urn, the w

orst-case computational and memory complexities scale quasi-linearly and linearl
y with the number of observed edges, making it efficient on sparse graphs, and r
equires only integer arithmetic. Key to our method is bits-back coding, which is
 used to sample edges and vertices without replacement from the edge-list in a w
ay that preserves the structure of the graph. Optimality is proven under a class
 of random graph models that are invariant to permutations of the edges and of v
ertices within an edge. Experiments indicate Random Edge Coding can achieve comp
etitive compression performance on real-world network datasets and scales to gra
phs with millions of nodes and edges.
****

ModelDiff: A Framework for Comparing Learning Algorithms
Harshay Shah, Sung Min Park, Andrew Ilyas, Aleksander Madry
We study the problem of (learning) algorithm comparison, where the goal is to fi
nd differences between models trained with two different learning algorithms. We
 begin by formalizing this goal as one of finding distinguishing feature transfo
rmations, i.e., input transformations that change the predictions of models trai
ned with one learning algorithm but not the other. We then present ModelDiff, a
method that leverages the datamodels framework (Ilyas et al., 2022) to compare l
earning algorithms based on how they use their training data. We demonstrate Mod
elDiff through three case studies, comparing models trained with/without data au
gmentation, with/without pre-training, and with different SGD hyperparameters.
****

Auxiliary Learning as an Asymmetric Bargaining Game
Aviv Shamsian, Aviv Navon, Neta Glazer, Kenji Kawaguchi, Gal Chechik, Ethan Feta
ya
Auxiliary learning is an effective method for enhancing the generalization capab
ilities of trained models, particularly when dealing with small datasets. Howeve
r, this approach may present several difficulties: (i) optimizing multiple objec
tives can be more challenging, and (ii) how to balance the auxiliary tasks to be
st assist the main task is unclear. In this work, we propose a novel approach, n
amed AuxiNash, for balancing tasks in auxiliary learning by formalizing the prob
lem as generalized bargaining game with asymmetric task bargaining power. Furthe
rmore, we describe an efficient procedure for learning the bargaining power of t
asks based on their contribution to the performance of the main task and derive
theoretical guarantees for its convergence. Finally, we evaluate AuxiNash on mul
tiple multi-task benchmarks and find that it consistently outperforms competing
methods.
****

Synthetic Prompting: Generating Chain-of-Thought Demonstrations for Large Langua
ge Models
Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, Weizhu Chen
Large language models can perform various reasoning tasks by using chain-of-thou
ght prompting, which guides them to find answers through step-by-step demonstrat
ions. However, the quality of the prompts depends on the demonstrations given to
 the models, and creating many of them by hand is costly. We introduce Synthetic
 prompting, a method that leverages a few handcrafted examples to prompt the mod
el to generate more examples by itself, and selects effective demonstrations to
elicit better reasoning. Our method alternates between a backward and forward pr
ocess to generate new examples. The backward process generates a question that m
atch a sampled reasoning chain, so that the question is solvable and clear. The
forward process produces a more detailed reasoning chain for the question, impro
ving the quality of the example. We evaluate our method on numerical, symbolic,
and algorithmic reasoning tasks, and show that it outperforms existing prompting
 techniques.
****

Complementary Attention for Multi-Agent Reinforcement Learning
Jianzhun Shao, Hongchang Zhang, Yun Qu, Chang Liu, Shuncheng He, Yuhang Jiang, X
iangyang Ji
In cooperative multi-agent reinforcement learning, centralized training with dec
entralized execution (CTDE) shows great promise for a trade-off between independ

ent Q-learning and joint action learning. However, vanilla CTDE methods assumed a fixed number of agents could hardly adapt to real-world scenarios where dynamic team compositions typically suffer from dramatically variant partial observability. Specifically, agents with extensive sight ranges are prone to be affected by trivial environmental substrates, dubbed the "distracted attention" issue; ones with limited observation can hardly sense their teammates, degrading the cooperation quality. In this paper, we propose Complementary Attention for Multi-Agent reinforcement learning (CAMA), which applies a divide-and-conquer strategy on input entities accompanied with the complementary attention of enhancement and replenishment. Concretely, to tackle the distracted attention issue, highly contributed entities' attention is enhanced by the execution-related representation extracted via action prediction with an inverse model. For better out-of-sight-range cooperation, the lowly contributed ones are compressed to brief messages with a conditional mutual information estimator. Our CAMA facilitates stable and sustainable teamwork, which is justified by the impressive results reported on the challenging StarCraftII, MPE, and Traffic Junction benchmarks.

****

Regularization-free Diffeomorphic Temporal Alignment Nets

Ron Shapira Weber, Oren Freifeld

In time-series analysis, nonlinear temporal misalignment is a major problem that forestalls even simple averaging. An effective learning-based solution for this problem is the Diffeomorphic Temporal Alignment Net (DTAN), that, by relying on a diffeomorphic temporal transformer net and the amortization of the joint-alignment task, eliminates drawbacks of traditional alignment methods. Unfortunately, existing DTAN formulations crucially depend on a regularization term whose optimal hyperparameters are dataset-specific and usually searched via a large number of experiments. Here we propose a regularization-free DTAN that obviates the need to perform such an expensive, and often impractical, search. Concretely, we propose a new well-behaved loss that we call the Inverse Consistency Averaging Error (ICAE), as well as a related new triplet loss. Extensive experiments on 128 UCR datasets show that the proposed method outperforms contemporary methods despite not using a regularization. Moreover, ICAE also gives rise to the first DTAN that supports variable-length signals. Our code is available at https://github.com/BGU-CS-VIL/RF-DTAN.

****

Toward Efficient Gradient-Based Value Estimation

Arsalan Sharifnassab, Richard S. Sutton

Gradient-based methods for value estimation in reinforcement learning have favorable stability properties, but they are typically much slower than Temporal Difference (TD) learning methods. We study the root causes of this slowness and show that Mean Square Bellman Error (MSBE) is an ill-conditioned loss function in the sense that its Hessian has large condition-number. To resolve the adverse effect of poor conditioning of MSBE on gradient based methods, we propose a low complexity batch-free proximal method that approximately follows the Gauss-Newton direction and is asymptotically robust to parameterization. Our main algorithm, called RANS, is efficient in the sense that it is significantly faster than the residual gradient methods while having almost the same computational complexity, and is competitive with TD on the classic problems that we tested.

****

Coin Sampling: Gradient-Based Bayesian Inference without Learning Rates

Louis Sharrock, Christopher Nemeth

In recent years, particle-based variational inference (ParVI) methods such as Stein variational gradient descent (SVGD) have grown in popularity as scalable methods for Bayesian inference. Unfortunately, the properties of such methods invariably depend on hyperparameters such as the learning rate, which must be carefully tuned by the practitioner in order to ensure convergence to the target measure at a suitable rate. In this paper, we introduce a suite of new particle-based methods for scalable Bayesian inference based on coin betting, which are entirely learning-rate free. We illustrate the performance of our approach on a range of numerical examples, including several high-dimensional models and datasets, de

monstrating comparable performance to other ParVI algorithms with no need to tun
e a learning rate.
****

On Kinetic Optimal Probability Paths for Generative Models
Neta Shaul, Ricky T. Q. Chen, Maximilian Nickel, Matthew Le, Yaron Lipman
Recent successful generative models are trained by fitting a neural network to a
n a-priori defined tractable probability density path taking noise to training e
xamples. In this paper we investigate the space of Gaussian probability paths, w
hich includes diffusion paths as an instance, and look for an optimal member in
some useful sense. In particular, minimizing the Kinetic Energy (KE) of a path i
s known to make particles' trajectories simple, hence easier to sample, and empi
rically improve performance in terms of likelihood of unseen data and sample gen
eration quality. We investigate Kinetic Optimal (KO) Gaussian paths and offer th
e following observations: (i) We show the KE takes a simplified form on the spac
e of Gaussian paths, where the data is incorporated only through a single, one d
imensional scalar function, called the data separation function. (ii) We charact
erize the KO solutions with a one dimensional ODE. (iii) We approximate data-dep
endent KO paths by approximating the data separation function and minimizing the
 KE. (iv) We prove that the data separation function converges to 1 in the gen
eral case of arbitrary normalized dataset consisting of n samples in d dimen
sion as $n/\sqrt{d}\rightarrow 0$. A consequence of this result is that the Cond
itional Optimal Transport (Cond-OT) path becomes kinetic optimal as $n/\sqrt{d}\
rightarrow 0$. We further support this theory with empirical experiments on Imag
eNet.
****

Sequential Changepoint Detection via Backward Confidence Sequences
Shubhanshu Shekhar, Aaditya Ramdas
We present a simple reduction from sequential estimation to sequential changepoi
nt detection (SCD). In short, suppose we are interested in detecting changepoint
s in some parameter or functional θ of the underlying distribution. We de
monstrate that if we can construct a confidence sequence (CS) for θ, then
 we can also successfully perform SCD for θ. This is accomplished by chec
king if two CSs — one forwards and the other backwards — ever fail to intersect.
 Since the literature on CSs has been rapidly evolving recently, the reduction p
rovided in this paper immediately solves several old and new change detection pr
oblems. Further, our "backward CS", constructed by reversing time, is new and po
tentially of independent interest. We provide strong nonasymptotic guarantees on
 the frequency of false alarms and detection delay, and demonstrate numerical ef
fectiveness on several problems.
****

Cold Analysis of Rao-Blackwellized Straight-Through Gumbel-Softmax Gradient Esti
mator
Alexander Shekhovtsov
Many problems in machine learning require an estimate of the gradient of an expe
ctation in discrete random variables with respect to the sampling distribution.
This work is motivated by the development of the Gumbel-Softmax family of estima
tors, which use a temperature-controlled relaxation of discrete variables. The s
tate-of-the art in this family, the Gumbel-Rao estimator uses an extra internal
sampling to reduce the variance, which may be costly. We analyze this estimator
and show that it possesses a zero temperature limit with a surprisingly simple c
losed form. The limit estimator, called ZGR, has favorable bias and variance pro
perties, it is easy to implement and computationally inexpensive. It decomposes
as the average of the straight through (ST) estimator and DARN estimator — two b
asic but not very well performing on their own estimators. We demonstrate that t
he simple ST–ZGR family of estimators practically dominates in the bias-variance
 tradeoffs the whole GR family while also outperforming SOTA unbiased estimators
.
****

Towards Understanding and Improving GFlowNet Training
Max W Shen, Emmanuel Bengio, Ehsan Hajiramezanali, Andreas Loukas, Kyunghyun Cho

, Tommaso Biancalani
Generative flow networks (GFlowNets) are a family of algorithms that learn a gen erative policy to sample discrete objects x with non-negative reward $R(x)$. L earning objectives guarantee the GFlowNet samples x from the target distributi on $p^*(x) \propto R(x)$ when loss is globally minimized over all states or traj ectories, but it is unclear how well they perform with practical limits on train ing resources. We introduce an efficient evaluation strategy to compare the lear ned sampling distribution to the target reward distribution. As flows can be und erdetermined given training data, we clarify the importance of learned flows to generalization and matching $p^*(x)$ in practice. We investigate how to learn be tter flows, and propose (i) prioritized replay training of high-reward x, (ii) relative edge flow policy parametrization, and (iii) a novel guided trajectory balance objective, and show how it can solve a substructure credit assignment pr oblem. We substantially improve sample efficiency on biochemical design tasks.
****

On Balancing Bias and Variance in Unsupervised Multi-Source-Free Domain Adaptati on
Maohao Shen, Yuheng Bu, Gregory W. Wornell
Due to privacy, storage, and other constraints, there is a growing need for unsu pervised domain adaptation techniques in machine learning that do not require ac cess to the data used to train a collection of source models. Existing methods f or multi-source-free domain adaptation (MSFDA) typically train a target model us ing pseudo-labeled data produced by the source models, which focus on improving the pseudo-labeling techniques or proposing new training objectives. Instead, we aim to analyze the fundamental limits of MSFDA. In particular, we develop an in formation-theoretic bound on the generalization error of the resulting target mo del, which illustrates an inherent bias-variance trade-off. We then provide insi ghts on how to balance this trade-off from three perspectives, including domain aggregation, selective pseudo-labeling, and joint feature alignment, which leads to the design of novel algorithms. Experiments on multiple datasets validate ou r theoretical analysis and demonstrate the state-of-art performance of the propo sed algorithm, especially on some of the most challenging datasets, including Of fice-Home and DomainNet.
****

On Penalty-based Bilevel Gradient Descent Method
Han Shen, Tianyi Chen
Bilevel optimization enjoys a wide range of applications in hyper-parameter opti mization, meta-learning and reinforcement learning. However, bilevel problems ar e difficult to solve and recent progress on scalable bilevel algorithms mainly f ocuses on bilevel optimization problems where the lower-level objective is eithe r strongly convex or unconstrained. In this work, we tackle the bilevel problem through the lens of the penalty method. We show that under certain conditions, t he penalty reformulation recovers the solutions of the original bilevel problem. Further, we propose the penalty-based bilevel gradient descent algorithm and es tablish its finite-time convergence for the constrained bilevel problem without lower-level strong convexity. The experimental results showcase the efficiency o f the proposed algorithm.
****

Non-autoregressive Conditional Diffusion Models for Time Series Prediction
Lifeng Shen, James Kwok
Recently, denoising diffusion models have led to significant breakthroughs in th e generation of images, audio and text. However, it is still an open question on how to adapt their strong modeling ability to model time series. In this paper, we propose TimeDiff, a non-autoregressive diffusion model that achieves high-qu ality time series prediction with the introduction of two novel conditioning mec hanisms: future mixup and autoregressive initialization. Similar to teacher forc ing, future mixup allows parts of the ground-truth future predictions for condit ioning, while autoregressive initialization helps better initialize the model wi th basic time series patterns such as short-term trends. Extensive experiments a re performed on nine real-world datasets. Results show that TimeDiff consistentl

y outperforms existing time series diffusion models, and also achieves the best overall performance across a variety of the existing strong baselines (including transformers and FiLM).
****

Cross-Modal Fine-Tuning: Align then Refine
Junhong Shen, Liam Li, Lucio M. Dery, Corey Staten, Mikhail Khodak, Graham Neubig, Ameet Talwalkar
Fine-tuning large-scale pretrained models has led to tremendous progress in well-studied modalities such as vision and NLP. However, similar gains have not been observed in many other modalities due to a lack of relevant pretrained models. In this work, we propose ORCA, a general cross-modal fine-tuning framework that extends the applicability of a single large-scale pretrained model to diverse modalities. ORCA adapts to a target task via an align-then-refine workflow: given the target input, ORCA first learns an embedding network that aligns the embedded feature distribution with the pretraining modality. The pretrained model is then fine-tuned on the embedded data to exploit the knowledge shared across modalities. Through extensive experiments, we show that ORCA obtains state-of-the-art results on 3 benchmarks containing over 60 datasets from 12 modalities, outperforming a wide range of hand-designed, AutoML, general-purpose, and task-specific cross-modal methods. We highlight the importance of data alignment via a series of ablation studies and exemplify ORCA’s utility in data-limited regimes.
****

Auxiliary Modality Learning with Generalized Curriculum Distillation
Yu Shen, Xijun Wang, Peng Gao, Ming Lin
Driven by the need from real-world applications, Auxiliary Modality Learning (AML) offers the possibility to utilize more information from auxiliary data in training, while only requiring data from one or fewer modalities in test, to save the overall computational cost and reduce the amount of input data for inferencing. In this work, we formally define “Auxiliary Modality Learning” (AML), systematically classify types of auxiliary modality (in visual computing) and architectures for AML, and analyze their performance. We also analyze the conditions under which AML works well from the optimization and data distribution perspectives. To guide various choices to achieve optimal performance using AML, we propose a novel method to assist in choosing the best auxiliary modality and estimating an upper bound performance before executing AML. In addition, we propose a new AML method using generalized curriculum distillation to enable more effective curriculum learning. Our method achieves the best performance compared to other SOTA methods.
****

TGRL: An Algorithm for Teacher Guided Reinforcement Learning
Idan Shenfeld, Zhang-Wei Hong, Aviv Tamar, Pulkit Agrawal
We consider solving sequential decision-making problems in the scenario where the agent has access to two supervision sources: $\textit{reward signal}$ and a $\textit{teacher}$ that can be queried to obtain a $\textit{good}$ action for any state encountered by the agent. Learning solely from rewards, or reinforcement learning, is data inefficient and may not learn high-reward policies in challenging scenarios involving sparse rewards or partial observability. On the other hand, learning from a teacher may sometimes be infeasible. For instance, the actions provided by a teacher with privileged information may be unlearnable by an agent with limited information (i.e., partial observability). In other scenarios, the teacher might be sub-optimal, and imitating their actions can limit the agent’s performance. To overcome these challenges, prior work proposed to jointly optimize imitation and reinforcement learning objectives but relied on heuristics and problem-specific hyper-parameter tuning to balance the two objectives. We introduce Teacher Guided Reinforcement Learning (TGRL), a principled approach to dynamically balance following the teacher’s guidance and leveraging RL. TGRL outperforms strong baselines across diverse domains without hyperparameter tuning.
****

FlexGen: High-Throughput Generative Inference of Large Language Models with a Single GPU

Ying Sheng, Lianmin Zheng, Binhang Yuan, Zhuohan Li, Max Ryabinin, Beidi Chen, P
ercy Liang, Christopher Re, Ion Stoica, Ce Zhang
The high computational and memory requirements of large language model (LLM) inf
erence make it feasible only with multiple high-end accelerators. Motivated by t
he emerging demand for latency-insensitive tasks with batched processing, this p
aper initiates the study of high-throughput LLM inference using limited resource
s, such as a single commodity GPU. We present FlexGen, a high-throughput generat
ion engine for running LLMs with limited GPU memory. FlexGen can be flexibly con
figured under various hardware resource constraints by aggregating memory and co
mputation from the GPU, CPU, and disk. By solving a linear programming problem,
it searches for efficient patterns to store and access tensors. FlexGen further
compresses the weights and the attention cache to 4 bits with negligible accurac
y loss. These techniques enable FlexGen to have a larger space of batch size cho
ices and thus significantly increase maximum throughput. As a result, when runni
ng OPT-175B on a single 16GB GPU, FlexGen achieves significantly higher throughp
ut compared to state-of-the-art offloading systems, reaching a generation throug
hput of 1 token/s for the first time with an effective batch size of 144. On the
 HELM benchmark, FlexGen can benchmark a 30B model with a 16GB GPU on 7 represen
tative sub-scenarios in 21 hours. The code is available at https://github.com/FM
Inference/FlexGen.
****
Improved Regret for Efficient Online Reinforcement Learning with Linear Function
 Approximation
Uri Sherman, Tomer Koren, Yishay Mansour
We study reinforcement learning with linear function approximation and adversari
ally changing cost functions, a setup that has mostly been considered under simp
lifying assumptions such as full information feedback or exploratory conditions.
 We present a computationally efficient policy optimization algorithm for the ch
allenging general setting of unknown dynamics and bandit feedback, featuring a c
ombination of mirror-descent and least squares policy evaluation in an auxiliary
 MDP used to compute exploration bonuses. Our algorithm obtains an $\widetilde O
(K^{6/7})$ regret bound, improving significantly over previous state-of-the-art
of $\widetilde O (K^{14/15})$ in this setting. In addition, we present a version
 of the same algorithm under the assumption a simulator of the environment is av
ailable to the learner (but otherwise no exploratory assumptions are made), and
prove it obtains state-of-the-art regret of $\widetilde O (K^{2/3})$.
****
Fundamental Limits of Two-layer Autoencoders, and Achieving Them with Gradient M
ethods
Aleksandr Shevchenko, Kevin Kögler, Hamed Hassani, Marco Mondelli
Autoencoders are a popular model in many branches of machine learning and lossy
data compression. However, their fundamental limits, the performance of gradient
 methods and the features learnt during optimization remain poorly understood, e
ven in the two-layer setting. In fact, earlier work has considered either linear
 autoencoders or specific training regimes (leading to vanishing or diverging co
mpression rates). Our paper addresses this gap by focusing on non-linear two-lay
er autoencoders trained in the challenging proportional regime in which the inpu
t dimension scales linearly with the size of the representation. Our results cha
racterize the minimizers of the population risk, and show that such minimizers a
re achieved by gradient methods; their structure is also unveiled, thus leading
to a concise description of the features obtained via training. For the special
case of a sign activation function, our analysis establishes the fundamental lim
its for the lossy compression of Gaussian sources via (shallow) autoencoders. Fi
nally, while the results are proved for Gaussian data, numerical simulations on
standard datasets display the universality of the theoretical predictions.
****
Large Language Models Can Be Easily Distracted by Irrelevant Context
Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, N
athanael Schärli, Denny Zhou
Large language models have achieved impressive performance on various natural la

nguage processing tasks. However, so far they have been evaluated primarily on b enchmarks where all information in the input context is relevant for solving the task. In this work, we investigate the distractibility of large language models , i.e., how the model prediction can be distracted by irrelevant context. In par ticular, we introduce Grade-School Math with Irrelevant Context (GSM-IC), an ari thmetic reasoning dataset with irrelevant information in the problem description . We use this benchmark to measure the distractibility of different prompting te chniques for large language models, and find that the model is easily distracted by irrelevant information. We also identify several approaches for mitigating t his deficiency, such as decoding with self-consistency and adding to the prompt an instruction that tells the language model to ignore the irrelevant informatio n.

****

Everyone’s Preference Changes Differently: A Weighted Multi-Interest Model For R etrieval

Hui Shi, Yupeng Gu, Yitong Zhou, Bo Zhao, Sicun Gao, Jishen Zhao

User embeddings (vectorized representations of a user) are essential in recommen dation systems. Numerous approaches have been proposed to construct a representa tion for the user in order to find similar items for retrieval tasks, and they h ave been proven effective in industrial recommendation systems. Recently people have discovered the power of using multiple embeddings to represent a user, with the hope that each embedding represents the user’s interest in a certain topic. With multi-interest representation, it’s important to model the user’s preferen ce over the different topics and how the preference changes with time. However, existing approaches either fail to estimate the user’s affinity to each interest or unreasonably assume every interest of every user fades at an equal rate with time, thus hurting the performance of candidate retrieval. In this paper, we pr opose the Multi-Interest Preference (MIP) model, an approach that not only produ ces multi-interest for users by using the user’s sequential engagement more effe ctively but also automatically learns a set of weights to represent the preferen ce over each embedding so that the candidates can be retrieved from each interes t proportionally. Extensive experiments have been done on various industrial-sca le datasets to demonstrate the effectiveness of our approach.

****

A Near-Optimal Algorithm for Safe Reinforcement Learning Under Instantaneous Har d Constraints

Ming Shi, Yingbin Liang, Ness Shroff

In many applications of Reinforcement Learning (RL), it is critically important that the algorithm performs safely, such that instantaneous hard constraints are satisfied at each step, and unsafe states and actions are avoided. However, exi sting algorithms for “safe” RL are often designed under constraints that either require expected cumulative costs to be bounded or assume all states are safe. T hus, such algorithms could violate instantaneous hard constraints and traverse u nsafe states (and actions) in practice. Hence, in this paper, we develop the fir st near-optimal safe RL algorithm for episodic Markov Decision Processes with un safe states and actions under instantaneous hard constraints and the linear mixt ure model. It achieves a regret $\tilde{O}(\frac{d H^3 \sqrt{d K}}{\Delta_c})$ t hat nearly matches the state-of-the-art regret in the setting with only unsafe a ctions and that in the unconstrained setting, and is safe at each step, where $d $ is the feature-mapping dimension, K is the number of episodes, H is the ep isode length, and Δ_c is a safety-related parameter. We also provide a lo wer bound $\tilde{\Omega}(\max\{d H \sqrt{K}, \frac{H}{\Delta_c^2}\})$, which in dicates that the dependency on Δ_c is necessary. Further, both our algori thm design and regret analysis involve several novel ideas, which may be of inde pendent interest.

****

Improving the Model Consistency of Decentralized Federated Learning

Yifan Shi, Li Shen, Kang Wei, Yan Sun, Bo Yuan, Xueqian Wang, Dacheng Tao

To mitigate the privacy leakages and communication burdens of Federated Learning (FL), decentralized FL (DFL) discards the central server and each client only c

ommunicates with its neighbors in a decentralized communication network. However, existing DFL suffers from high inconsistency among local clients, which results in severe distribution shift and inferior performance compared with centralized FL (CFL), especially on heterogeneous data or sparse communication topologies. To alleviate this issue, we propose two DFL algorithms named DFedSAM and DFedSAM-MGS to improve the performance of DFL. Specifically, DFedSAM leverages gradient perturbation to generate local flat models via Sharpness Aware Minimization (SAM), which searches for models with uniformly low loss values. DFedSAM-MGS further boosts DFedSAM by adopting Multiple Gossip Steps (MGS) for better model consistency, which accelerates the aggregation of local flat models and better balances communication complexity and generalization. Theoretically, we present improved convergence rates $\small \mathcal{O}\big(\frac{1}{\sqrt{KT}}+\frac{1}{T}+\frac{1}{K^{1/2}T^{3/2}(1-\lambda)^2}\big)$ and $\small \mathcal{O}\big(\frac{1}{\sqrt{KT}}+\frac{1}{T}+\frac{\lambda^Q+1}{K^{1/2}T^{3/2}(1-\lambda^Q)^2}\big)$ in non-convex setting for DFedSAM and DFedSAM-MGS, respectively, where $1-\lambda$ is the spectral gap of gossip matrix and Q is the number of MGS. Empirically, our methods can achieve competitive performance compared with CFL methods and outperform existing DFL methods.

****

UPop: Unified and Progressive Pruning for Compressing Vision-Language Transformers

Dachuan Shi, Chaofan Tao, Ying Jin, Zhendong Yang, Chun Yuan, Jiaqi Wang

Real-world data contains a vast amount of multimodal information, among which vision and language are the two most representative modalities. Moreover, increasingly heavier models, e.g., Transformers, have attracted the attention of researchers to model compression. However, how to compress multimodal models, especially vison-language Transformers, is still under-explored. This paper proposes the Unified and Progressive Pruning (UPop) as a universal vison-language Transformer compression framework, which incorporates 1) unifiedly searching multimodal subnets in a continuous optimization space from the original model, which enables automatic assignment of pruning ratios among compressible modalities and structures; 2) progressively searching and retraining the subnet, which maintains convergence between the search and retrain to attain higher compression ratios. Experiments on various tasks, datasets, and model architectures demonstrate the effectiveness and versatility of the proposed UPop framework. The code is available at https://github.com/sdc17/UPop.

****

Sequence Modeling with Multiresolution Convolutional Memory

Jiaxin Shi, Ke Alexander Wang, Emily Fox

Efficiently capturing the long-range patterns in sequential data sources salient to a given task—such as classification and generative modeling—poses a fundamental challenge. Popular approaches in the space tradeoff between the memory burden of brute-force enumeration and comparison, as in transformers, the computational burden of complicated sequential dependencies, as in recurrent neural networks, or the parameter burden of convolutional networks with many or large filters. We instead take inspiration from wavelet-based multiresolution analysis to define a new building block for sequence modeling, which we call a MultiresLayer. The key component of our model is the multiresolution convolution, capturing multiscale trends in the input sequence. Our MultiresConv can be implemented with shared filters across a dilated causal convolution tree. Thus it garners the computational advantages of convolutional networks and the principled theoretical motivation of wavelet decompositions. Our MultiresLayer is straightforward to implement, requires significantly fewer parameters, and maintains at most a $O(N \log N)$ memory footprint for a length N sequence. Yet, by stacking such layers, our model yields state-of-the-art performance on a number of sequence classification and autoregressive density estimation tasks using CIFAR-10, ListOps, and PTB-XL datasets.

****

Statistical Inference on Multi-armed Bandits with Delayed Feedback

Lei Shi, Jingshen Wang, Tianhao Wu

Multi armed bandit (MAB) algorithms have been increasingly used to complement or integrate with A/B tests and randomized clinical trials in e-commerce, healthcare, and policymaking. Recent developments incorporate possible delayed feedback. While existing MAB literature often focuses on maximizing the expected cumulative reward outcomes (or, equivalently, regret minimization), few efforts have been devoted to establish valid statistical inference approaches to quantify the uncertainty of learned policies. We attempt to fill this gap by providing a unified statistical inference framework for policy evaluation where a target policy is allowed to differ from the data collecting policy, and our framework allows delay to be associated with the treatment arms. We present an adaptively weighted estimator that on one hand incorporates the arm-dependent delaying mechanism to achieve consistency, and on the other hand mitigates the variance inflation across stages due to vanishing sampling probability. In particular, our estimator does not critically depend on the ability to estimate the unknown delay mechanism. Under appropriate conditions, we prove that our estimator converges to a normal distribution as the number of time points goes to infinity, which provides guarantees for large-sample statistical inference. We illustrate the finite-sample performance of our approach through Monte Carlo experiments.

****

Provably Efficient Offline Reinforcement Learning with Perturbed Data Sources

Chengshuai Shi, Wei Xiong, Cong Shen, Jing Yang

Existing theoretical studies on offline reinforcement learning (RL) mostly consider a dataset sampled directly from the target task. In practice, however, data often come from several heterogeneous but related sources. Motivated by this gap, this work aims at rigorously understanding offline RL with multiple datasets that are collected from randomly perturbed versions of the target task instead of from itself. An information-theoretic lower bound is derived, which reveals a necessary requirement on the number of involved sources in addition to that on the number of data samples. Then, a novel HetPEVI algorithm is proposed, which simultaneously considers the sample uncertainties from a finite number of data samples per data source and the source uncertainties due to a finite number of available data sources. Theoretical analyses demonstrate that HetPEVI can solve the target task as long as the data sources collectively provide a good data coverage. Moreover, HetPEVI is demonstrated to be optimal up to a polynomial factor of the horizon length. Finally, the study is extended to offline Markov games and offline robust RL, which demonstrates the generality of the proposed designs and theoretical analyses.

****

On the Complexity of Bayesian Generalization

Yu-Zhe Shi, Manjie Xu, John E. Hopcroft, Kun He, Joshua B. Tenenbaum, Song-Chun Zhu, Ying Nian Wu, Wenjuan Han, Yixin Zhu

We examine concept generalization at a large scale in the natural visual spectrum. Established computational modes (i.e., rule-based or similarity-based) are primarily studied isolated, focusing on confined and abstract problem spaces. In this work, we study these two modes when the problem space scales up and when the complexity of concepts becomes diverse. At the representational level, we investigate how the complexity varies when a visual concept is mapped to the representation space. Prior literature has shown that two types of complexities (Griffiths & Tenenbaum, 2003) build an inverted-U relation (Donderi, 2006; Sun & Firestone, 2021). Leveraging Representativeness of Attribute (RoA), we computationally confirm: Models use attributes with high RoA to describe visual concepts, and the description length falls in an inverted-U relation with the increment in visual complexity. At the computational level, we examine how the complexity of representation affects the shift between the rule- and similarity-based generalization. We hypothesize that category-conditioned visual modeling estimates the co-occurrence frequency between visual and categorical attributes, thus potentially serving as the prior for the natural visual world. Experimental results show that representations with relatively high subjective complexity outperform those with relatively low subjective complexity in rule-based generalization, while the trend is the opposite in similarity-based generalization.

****
Understanding and Generalizing Contrastive Learning from the Inverse Optimal Transport Perspective
Liangliang Shi, Gu Zhang, Haoyu Zhen, Jintao Fan, Junchi Yan
Previous research on contrastive learning (CL) has primarily focused on pairwise views to learn representations by attracting positive samples and repelling negative ones. In this work, we aim to understand and generalize CL from a point set matching perspective, instead of the comparison between two points. Specifically, we formulate CL as a form of inverse optimal transport (IOT), which involves a bilevel optimization procedure for learning where the outter minimization aims to learn the representations and the inner is to learn the coupling (i.e. the probability of matching matrix) between the point sets. Specifically, by adjusting the relaxation degree of constraints in the inner minimization, we obtain three contrastive losses and show that the dominant contrastive loss in literature InfoNCE falls into one of these losses. This reveals a new and more general algorithmic framework for CL. Additionally, the soft matching scheme in IOT induces a uniformity penalty to enhance representation learning which is akin to the CL's uniformity. Results on vision benchmarks show the effectiveness of our derived loss family and the new uniformity term.
****
Long Horizon Temperature Scaling
Andy Shih, Dorsa Sadigh, Stefano Ermon
Temperature scaling is a popular technique for tuning the sharpness of a model distribution. It is used extensively for sampling likely generations and calibrating model uncertainty, and even features as a controllable parameter to many large language models in deployment. However, autoregressive models rely on myopic temperature scaling that greedily optimizes the next token. To address this, we propose Long Horizon Temperature Scaling (LHTS), a novel approach for sampling from temperature-scaled joint distributions. LHTS is compatible with all likelihood-based models, and optimizes for the long-horizon likelihood of samples. We derive a temperature-dependent LHTS objective, and show that fine-tuning a model on a range of temperatures produces a single model capable of generation with a controllable long-horizon temperature parameter. We experiment with LHTS on image diffusion models and character/language autoregressive models, demonstrating its advantages over myopic temperature scaling in likelihood and sample quality, and showing improvements in accuracy of a multiple choice analogy by 10%.
****
Gradient Descent in Neural Networks as Sequential Learning in Reproducing Kernel Banach Space
Alistair Shilton, Sunil Gupta, Santu Rana, Svetha Venkatesh
The study of Neural Tangent Kernels (NTKs) has provided much needed insight into convergence and generalization properties of neural networks in the over-parametrized (wide) limit by approximating the network using a first-order Taylor expansion with respect to its weights in the neighborhood of their initialization values. This allows neural network training to be analyzed from the perspective of reproducing kernel Hilbert spaces (RKHS), which is informative in the over-parametrized regime, but a poor approximation for narrower networks as the weights change more during training. Our goal is to extend beyond the limits of NTK toward a more general theory. We construct an exact power-series representation of the neural network in a finite neighborhood of the initial weights as an inner product of two feature maps, respectively from data and weight-step space, to feature space, allowing neural network training to be analyzed from the perspective of reproducing kernel Banach space (RKBS). We prove that, regardless of width, the training sequence produced by gradient descent can be exactly replicated by regularized sequential learning in RKBS. Using this, we present novel bound on uniform convergence where the iterations count and learning rate play a central role, giving new theoretical insight into neural network training.
****
SNeRL: Semantic-aware Neural Radiance Fields for Reinforcement Learning
Dongseok Shim, Seungjae Lee, H. Jin Kim

As previous representations for reinforcement learning cannot effectively incorporate a human-intuitive understanding of the 3D environment, they usually suffer from sub-optimal performances. In this paper, we present Semantic-aware Neural Radiance Fields for Reinforcement Learning (SNeRL), which jointly optimizes semantic-aware neural radiance fields (NeRF) with a convolutional encoder to learn 3D-aware neural implicit representation from multi-view images. We introduce 3D semantic and distilled feature fields in parallel to the RGB radiance fields in NeRF to learn semantic and object-centric representation for reinforcement learning. SNeRL outperforms not only previous pixel-based representations but also recent 3D-aware representations both in model-free and model-based reinforcement learning.

****

A Closer Look at the Intervention Procedure of Concept Bottleneck Models

Sungbin Shin, Yohan Jo, Sungsoo Ahn, Namhoon Lee

Concept bottleneck models (CBMs) are a class of interpretable neural network models that predict the target response of a given input based on its high-level concepts. Unlike the standard end-to-end models, CBMs enable domain experts to intervene on the predicted concepts and rectify any mistakes at test time, so that more accurate task predictions can be made at the end. While such intervenability provides a powerful avenue of control, many aspects of the intervention procedure remain rather unexplored. In this work, we develop various ways of selecting intervening concepts to improve the intervention effectiveness and conduct an array of in-depth analyses as to how they evolve under different circumstances. Specifically, we find that an informed intervention strategy can reduce the task error more than ten times compared to the current baseline under the same amount of intervention counts in realistic settings, and yet, this can vary quite significantly when taking into account different intervention granularity. We verify our findings through comprehensive evaluations, not only on the standard real datasets, but also on synthetic datasets that we generate based on a set of different causal graphs. We further discover some major pitfalls of the current practices which, without a proper addressing, raise concerns on reliability and fairness of the intervention procedure.

****

MetricGAN-OKD: Multi-Metric Optimization of MetricGAN via Online Knowledge Distillation for Speech Enhancement

Wooseok Shin, Byung Hoon Lee, Jin Sob Kim, Hyun Joon Park, Sung Won Han

In speech enhancement, MetricGAN-based approaches reduce the discrepancy between the L_p loss and evaluation metrics by utilizing a non-differentiable evaluation metric as the objective function. However, optimizing multiple metrics simultaneously remains challenging owing to the problem of confusing gradient directions. In this paper, we propose an effective multi-metric optimization method in MetricGAN via online knowledge distillation—MetricGAN-OKD. MetricGAN-OKD, which consists of multiple generators and target metrics, related by a one-to-one correspondence, enables generators to learn with respect to a single metric reliably while improving performance with respect to other metrics by mimicking other generators. Experimental results on speech enhancement and listening enhancement tasks reveal that the proposed method significantly improves performance in terms of multiple metrics compared to existing multi-metric optimization methods. Further, the good performance of MetricGAN-OKD is explained in terms of network generalizability and correlation between metrics.

****

Improved Learning-Augmented Algorithms for the Multi-Option Ski Rental Problem via Best-Possible Competitive Analysis

Yongho Shin, Changyeol Lee, Gukryeol Lee, Hyung-Chan An

In this paper, we present improved learning-augmented algorithms for the multi-option ski rental problem. Learning-augmented algorithms take ML predictions as an added part of the input and incorporates these predictions in solving the given problem. Due to their unique strength that combines the power of ML predictions with rigorous performance guarantees, they have been extensively studied in the context of online optimization problems. Even though ski rental problems are o

ne of the canonical problems in the field of online optimization, only determini stic algorithms were previously known for multi-option ski rental, with or witho ut learning augmentation. We present the first randomized learning-augmented alg orithm for this problem, surpassing previous performance guarantees given by det erministic algorithms. Our learning-augmented algorithm is based on a new, prova bly best-possible randomized competitive algorithm for the problem. Our results are further complemented by lower bounds for deterministic and randomized algori thms, and computational experiments evaluating our algorithms' performance impro vements.

****

One-shot Imitation in a Non-Stationary Environment via Multi-Modal Skill

Sangwoo Shin, Daehee Lee, Minjong Yoo, Woo Kyung Kim, Honguk Woo

One-shot imitation is to learn a new task from a single demonstration, yet it is a challenging problem to adopt it for complex tasks with the high domain divers ity inherent in a non-stationary environment. To tackle the problem, we explore the compositionality of complex tasks, and present a novel skill-based imitation learning framework enabling one-shot imitation and zero-shot adaptation; from a single demonstration for a complex unseen task, a semantic skill sequence is in ferred and then each skill in the sequence is converted into an action sequence optimized for environmental hidden dynamics that can vary over time. Specificall y, we leverage a vision-language model to learn a semantic skill set from offlin e video datasets, where each skill is represented on the vision-language embeddi ng space, and adapt meta-learning with dynamics inference to enable zero-shot sk ill adaptation. We evaluate our framework with various one-shot imitation scenar ios for extended multi-stage Meta-world tasks, showing its superiority in learni ng complex tasks, generalizing to dynamics changes, and extending to different d emonstration conditions and modalities, compared to other baselines.

****

Context Consistency Regularization for Label Sparsity in Time Series

Yooju Shin, Susik Yoon, Hwanjun Song, Dongmin Park, Byunghyun Kim, Jae-Gil Lee, Byung Suk Lee

Labels are typically sparse in real-world time series due to the high annotation cost. Recently, consistency regularization techniques have been used to generat e artificial labels from unlabeled augmented instances. To fully exploit the seq uential characteristic of time series in consistency regularization, we propose a novel method of data augmentation called context-attached augmentation, which adds preceding and succeeding instances to a target instance to form its augment ed instance. Unlike the existing augmentation techniques that modify a target in stance by directly perturbing its attributes, the context-attached augmentation generates instances augmented with varying contexts while maintaining the target instance. Based on our augmentation method, we propose a context consistency re gularization framework, which first adds different contexts to a target instance sampled from a given time series and then shares unitary reliability-based cros s-window labels across the augmented instances to maintain consistency. We demon strate that the proposed framework outperforms the existing state-of-the-art con sistency regularization frameworks through comprehensive experiments on real-wor ld time-series datasets.

****

Generative Causal Representation Learning for Out-of-Distribution Motion Forecas ting

Shayan Shirahmad Gale Bagi, Zahra Gharaee, Oliver Schulte, Mark Crowley

Conventional supervised learning methods typically assume i.i.d samples and are found to be sensitive to out-of-distribution (OOD) data. We propose Generative C ausal Representation Learning (GCRL) which leverages causality to facilitate kno wledge transfer under distribution shifts. While we evaluate the effectiveness o f our proposed method in human trajectory prediction models, GCRL can be applied to other domains as well. First, we propose a novel causal model that explains the generative factors in motion forecasting datasets using features that are co mmon across all environments and with features that are specific to each environ ment. Selection variables are used to determine which parts of the model can be

directly transferred to a new environment without fine-tuning. Second, we propos e an end-to-end variational learning paradigm to learn the causal mechanisms tha t generate observations from features. GCRL is supported by strong theoretical r esults that imply identifiability of the causal model under certain assumptions. Experimental results on synthetic and real-world motion forecasting datasets sh ow the robustness and effectiveness of our proposed method for knowledge transfe r under zero-shot and low-shot settings by substantially outperforming the prior motion forecasting models on out-of-distribution prediction.
****

Exphormer: Sparse Transformers for Graphs
Hamed Shirzad, Ameya Velingker, Balaji Venkatachalam, Danica J. Sutherland, Ali Kemal Sinop
Graph transformers have emerged as a promising architecture for a variety of gra ph learning and representation tasks. Despite their successes, though, it remain s challenging to scale graph transformers to large graphs while maintaining accu racy competitive with message-passing networks. In this paper, we introduce Exph ormer, a framework for building powerful and scalable graph transformers. Exphor mer consists of a sparse attention mechanism based on two mechanisms: virtual gl obal nodes and expander graphs, whose mathematical characteristics, such as spec tral expansion, pseduorandomness, and sparsity, yield graph transformers with co mplexity only linear in the size of the graph, while allowing us to prove desira ble theoretical properties of the resulting transformer models. We show that inc orporating Exphormer into the recently-proposed GraphGPS framework produces mode ls with competitive empirical results on a wide variety of graph datasets, inclu ding state-of-the-art results on three datasets. We also show that Exphormer can scale to datasets on larger graphs than shown in previous graph transformer arc hitectures.
****

Synthetic data for model selection
Alon Shoshan, Nadav Bhonker, Igor Kviatkovsky, Matan Fintz, Gerard Medioni
Recent breakthroughs in synthetic data generation approaches made it possible to produce highly photorealistic images which are hardly distinguishable from real ones. Furthermore, synthetic generation pipelines have the potential to generat e an unlimited number of images. The combination of high photorealism and scale turn synthetic data into a promising candidate for improving various machine lea rning (ML) pipelines. Thus far, a large body of research in this field has focus ed on using synthetic images for training, by augmenting and enlarging training data. In contrast to using synthetic data for training, in this work we explore whether synthetic data can be beneficial for model selection. Considering the ta sk of image classification, we demonstrate that when data is scarce, synthetic d ata can be used to replace the held out validation set, thus allowing to train o n a larger dataset. We also introduce a novel method to calibrate the synthetic error estimation to fit that of the real domain. We show that such calibration s ignificantly improves the usefulness of synthetic data for model selection.
****

Probabilistic Attention-to-Influence Neural Models for Event Sequences
Xiao Shou, Debarun Bhattacharjya, Tian Gao, Dharmashankar Subramanian, Oktie Has sanzadeh, Kristin Bennett
Discovering knowledge about which types of events influence others, using datase ts of event sequences without time stamps, has several practical applications. W hile neural sequence models are able to capture complex and potentially long-ran ge historical dependencies, they often lack the interpretability of simpler mode ls for event sequence dynamics. We provide a novel neural framework in such a se tting - a probabilistic attention-to-influence neural model - which not only cap tures complex instance-wise interactions between events but also learns influenc ers for each event type of interest. Given event sequence data and a prior distr ibution on type-wise influence, we efficiently learn an approximate posterior fo r type-wise influence by an attention-to-influence transformation using variatio nal inference. Our method subsequently models the conditional likelihood of sequ ences by sampling the above posterior to focus attention on influencing event ty

pes. We motivate our general framework and show improved performance in experime nts compared to existing baselines on synthetic data as well as real-world bench marks, for tasks involving prediction and influencing set identification.
****

Causal Bounds in Quasi-Markovian Graphs
Madhumitha Shridharan, Garud Iyengar
We consider the problem of computing bounds for causal queries on quasi-Markovia n graphs with unobserved confounders and discrete valued observed variables, whe re identifiability does not hold. Existing non-parametric approaches for computi ng such bounds use multilinear programming (MP) formulations that are often intr actable for existing solvers when the degree of the polynomial objective is grea ter than two. Hence, one often has to resort to either fast approximate heuristi cs which are not guaranteed to contain the true query value, or more accurate bu t computationally intensive procedures. We show how to construct an equivalent M P with a polynomial objective of lower degree. In particular, the degree of the objective in the new MP is equal to only the number of C-components that are int ervened upon, instead of the total number of C-components. As a result, we can c ompute exact bounds for significantly larger causal inference problems as compar ed to what is possible using existing techniques. We also propose a very efficie nt Frank-Wolfe heuristic that produces very high quality bounds, and scales to l arge multilinear problems of higher degree.
****

Repository-Level Prompt Generation for Large Language Models of Code
Disha Shrivastava, Hugo Larochelle, Daniel Tarlow
With the success of large language models (LLMs) of code and their use as code a ssistants (e.g. Codex used in GitHub Copilot), techniques for introducing domain -specific knowledge in the prompt design process become important. In this work, we propose a framework called Repo-Level Prompt Generator that learns to genera te example-specific prompts using prompt proposals. The prompt proposals take co ntext from the entire repository, thereby incorporating both the structure of th e repository and the context from other relevant files (e.g. imports, parent cla ss files). Our technique doesn’t require any access to the weights of the LLM, m aking it applicable in cases where we only have black-box access to the LLM. We conduct experiments on the task of single-line code auto-completion using code r epositories taken from Google Code archives. We demonstrate that an oracle const ructed from our prompt proposals gives a relative improvement of 36% over Codex, showing the quality of these proposals. Further, we show that when we train a m odel to predict a prompt proposal, we can achieve significant performance gains over Codex and other baselines. We release our code, data, and trained checkpoin ts at: https://github.com/shrivastavadisha/repo_level_prompt_generation.
****

CLIPood: Generalizing CLIP to Out-of-Distributions
Yang Shu, Xingzhuo Guo, Jialong Wu, Ximei Wang, Jianmin Wang, Mingsheng Long
Out-of-distribution (OOD) generalization, where the model needs to handle distri bution shifts from training, is a major challenge of machine learning. Contrasti ve language-image pre-training (CLIP) models have shown impressive zero-shot abi lity, but the further adaptation of CLIP on downstream tasks undesirably degrade s OOD performances. This paper aims at generalizing CLIP to out-of-distribution test data on downstream tasks. We propose CLIPood, a fine-tuning method that can adapt CLIP models to OOD situations where both domain shifts and open classes m ay occur on the unseen test data. To exploit the semantic relations between clas ses from the text modality, CLIPood introduces a new training objective, margin metric softmax (MMS), with class adaptive margins for fine-tuning. To incorporat e both pre-trained zero-shot model and fine-tuned task-adaptive model, CLIPood l everages a new optimization strategy, Beta moving average (BMA), to maintain a t emporal ensemble weighted by Beta distribution. Experiments on diverse datasets with different OOD scenarios show that CLIPood consistently outperforms existing generalization techniques.
****

Semi-Autoregressive Energy Flows: Exploring Likelihood-Free Training of Normaliz

ing Flows
Phillip Si, Zeyi Chen, Subham Sekhar Sahoo, Yair Schiff, Volodymyr Kuleshov
Training normalizing flow generative models can be challenging due to the need t o calculate computationally expensive determinants of Jacobians. This paper stud ies the likelihood-free training of flows and proposes the energy objective, an alternative sample-based loss based on proper scoring rules. The energy objectiv e is determinant-free and supports flexible model architectures that are not eas ily compatible with maximum likelihood training, including semi-autoregressive e nergy flows, a novel model family that interpolates between fully autoregressive and non-autoregressive models. Energy flows feature competitive sample quality, posterior inference, and generation speed relative to likelihood-based flows; t his performance is decorrelated from the quality of log-likelihood estimates, wh ich are generally very poor. Our findings question the use of maximum likelihood as an objective or a metric, and contribute to a scientific study of its role i n generative modeling. Code is available at https://github.com/ps789/SAEF.
****
Unearthing InSights into Mars: Unsupervised Source Separation with Limited Data
Ali Siahkoohi, Rudy Morel, Maarten V. De Hoop, Erwan Allys, Gregory Sainton, Tai chi Kawamura
Source separation involves the ill-posed problem of retrieving a set of source s ignals that have been observed through a mixing operator. Solving this problem r equires prior knowledge, which is commonly incorporated by imposing regularity c onditions on the source signals, or implicitly learned through supervised or uns upervised methods from existing data. While data-driven methods have shown great promise in source separation, they often require large amounts of data, which r arely exists in planetary space missions. To address this challenge, we propose an unsupervised source separation scheme for domains with limited data access th at involves solving an optimization problem in the wavelet scattering covariance representation space—an interpretable, low-dimensional representation of statio nary processes. We present a real-data example in which we remove transient, the rmally-induced microtilts—known as glitches—from data recorded by a seismometer during NASA’s InSight mission on Mars. Thanks to the wavelet scattering covarian ces’ ability to capture non-Gaussian properties of stochastic processes, we are able to separate glitches using only a few glitch-free data snippets.
****
Quantitative Universal Approximation Bounds for Deep Belief Networks
Julian Sieber, Johann Gehringer
We show that deep belief networks with binary hidden units can approximate any m ultivariate probability density under very mild integrability requirements on th e parental density of the visible nodes. The approximation is measured in the $L ^q$-norm for $q\in[1,\infty]$ ($q=\infty$ corresponding to the supremum norm) an d in Kullback-Leibler divergence. Furthermore, we establish sharp quantitative b ounds on the approximation error in terms of the number of hidden units.
****
Pricing Experimental Design: Causal Effect, Expected Revenue and Tail Risk
David Simchi-Levi, Chonghuan Wang
When launching a new product, historical sales data is often not available, leav ing price as a crucial experimental instrument for sellers to gauge market respo nse. When designing pricing experiments, there are three fundamental objectives: estimating the causal effect of price (i.e., price elasticity), maximizing the expected revenue through the experiment, and controlling the tail risk suffering from a very huge loss. In this paper, we reveal the relationship among such thr ee objectives. Under a linear structural model, we investigate the trade-offs be tween causal inference and expected revenue maximization, as well as between exp ected revenue maximization and tail risk control. Furthermore, we propose an opt imal pricing experimental design, which can flexibly adapt to different desired levels of trade-offs. Through the optimal design, we also explore the relationsh ip between causal inference and tail risk control.
****
Statistical Learning under Heterogeneous Distribution Shift

Max Simchowitz, Anurag Ajay, Pulkit Agrawal, Akshay Krishnamurthy
This paper studies the prediction of a target $\mathbf{z}$ from a pair of random variables $(\mathbf{x},\mathbf{y})$, where the ground-truth predictor is additive $\mathbb{E}[\mathbf{z} \mid \mathbf{x},\mathbf{y}] = f_\star(\mathbf{x}) +g_{\star}(\mathbf{y})$. We study the performance of empirical risk minimization (ERM) over functions $f+g$, $f \in \mathcal{F}$ and $g \in \mathcal{G}$, fit on a given training distribution, but evaluated on a test distribution which exhibits covariate shift. We show that, when the class $\mathcal{F}$ is "simpler" than $\mathcal{G}$ (measured, e.g., in terms of its metric entropy), our predictor is more resilient to heterogeneous covariate shifts in which the shift in $\mathbf{x}$ is much greater than that in $\mathbf{y}$. These results rely on a novel Hölder style inequality for the Dudley integral which may be of independent interest. Moreover, we corroborate our theoretical findings with experiments demonstrating improved resilience to shifts in "simpler" features across numerous domains.
****
On the Stepwise Nature of Self-Supervised Learning
James B Simon, Maksis Knutins, Liu Ziyin, Daniel Geisz, Abraham J Fetterman, Joshua Albrecht
We present a simple picture of the training process of self-supervised learning methods with dual deep networks. In our picture, these methods learn their high-dimensional embeddings one dimension at a time in a sequence of discrete, well-separated steps. We arrive at this picture via the study of a linear toy model of Barlow Twins, applicable to the case in which the trained network is infinitely wide. We solve the training dynamics of our toy model from small initialization, finding that the model learns the top eigenmodes of a certain contrastive kernel in a discrete, stepwise fashion, and find a closed-form expression for the final learned representations. Remarkably, we see the same stepwise learning phenomenon when training deep ResNets using the Barlow Twins, SimCLR, and VICReg losses. This stepwise picture partially demystifies the process of self-supervised training.
****
Hindsight Learning for MDPs with Exogenous Inputs
Sean R. Sinclair, Felipe Vieira Frujeri, Ching-An Cheng, Luke Marshall, Hugo De Oliveira Barbalho, Jingling Li, Jennifer Neville, Ishai Menache, Adith Swaminathan
Many resource management problems require sequential decision-making under uncertainty, where the only uncertainty affecting the decision outcomes are exogenous variables outside the control of the decision-maker. We model these problems as Exo-MDPs (Markov Decision Processes with Exogenous Inputs) and design a class of data-efficient algorithms for them termed Hindsight Learning (HL). Our HL algorithms achieve data efficiency by leveraging a key insight: having samples of the exogenous variables, past decisions can be revisited in hindsight to infer counterfactual consequences that can accelerate policy improvements. We compare HL against classic baselines in the multi-secretary and airline revenue management problems. We also scale our algorithms to a business-critical cloud resource management problem – allocating Virtual Machines (VMs) to physical machines, and simulate their performance with real datasets from a large public cloud provider. We find that HL algorithms outperform domain-specific heuristics, as well as state-of-the-art reinforcement learning methods.
****
Text-To-4D Dynamic Scene Generation
Uriel Singer, Shelly Sheynin, Adam Polyak, Oron Ashual, Iurii Makarov, Filippos Kokkinos, Naman Goyal, Andrea Vedaldi, Devi Parikh, Justin Johnson, Yaniv Taigman
We present MAV3D (Make-A-Video3D), a method for generating three-dimensional dynamic scenes from text descriptions. Our approach uses a 4D dynamic Neural Radiance Field (NeRF), which is optimized for scene appearance, density, and motion consistency by querying a Text-to-Video (T2V) diffusion-based model. The dynamic video output generated from the provided text can be viewed from any camera location and angle, and can be composited into any 3D environment. MAV3D does not req

uire any 3D or 4D data and the T2V model is trained only on Text-Image pairs and unlabeled videos. We demonstrate the effectiveness of our approach using comprehensive quantitative and qualitative experiments and show an improvement over previously established internal baselines. To the best of our knowledge, our method is the first to generate 3D dynamic scenes given a text description. Generated samples can be viewed at make-a-video3d.github.io
****

The Hessian perspective into the Nature of Convolutional Neural Networks
Sidak Pal Singh, Thomas Hofmann, Bernhard Schölkopf
While Convolutional Neural Networks (CNNs) have long been investigated and applied, as well as theorized, we aim to provide a slightly different perspective into their nature — through the perspective of their Hessian maps. The reason is that the loss Hessian captures the pairwise interaction of parameters and therefore forms a natural ground to probe how the architectural aspects of CNNs get manifested in their structure and properties. We develop a framework relying on Toeplitz representation of CNNs, and then utilize it to reveal the Hessian structure and, in particular, its rank. We prove tight upper bounds (with linear activations), which closely follow the empirical trend of the Hessian rank and in practice also hold for more general settings. Overall, our work generalizes and further establishes the key insight that the Hessian rank grows as the square root of the number of parameters, even in CNNs.
****

When do Minimax-fair Learning and Empirical Risk Minimization Coincide?
Harvineet Singh, Matthäus Kleindessner, Volkan Cevher, Rumi Chunara, Chris Russell
Minimax-fair machine learning minimizes the error for the worst-off group. However, empirical evidence suggests that when sophisticated models are trained with standard empirical risk minimization (ERM), they often have the same performance on the worst-off group as a minimax-trained model. Our work makes this counter-intuitive observation concrete. We prove that if the hypothesis class is sufficiently expressive and the group information is recoverable from the features, ERM and minimax-fairness learning formulations indeed have the same performance on the worst-off group. We provide additional empirical evidence of how this observation holds on a wide range of datasets and hypothesis classes. Since ERM is fundamentally easier than minimax optimization, our findings have implications on the practice of fair machine learning.
****

Differentiable Simulations for Enhanced Sampling of Rare Events
Martin Sipka, Johannes C. B. Dietschreit, Lukáš Grajciar, Rafael Gomez-Bombarelli
Simulating rare events, such as the transformation of a reactant into a product in a chemical reaction typically requires enhanced sampling techniques that rely on heuristically chosen collective variables (CVs). We propose using differentiable simulations (DiffSim) for the discovery and enhanced sampling of chemical transformations without a need to resort to preselected CVs, using only a distance metric. Reaction path discovery and estimation of the biasing potential that enhances the sampling are merged into a single end-to-end problem that is solved by path-integral optimization. This is achieved by introducing multiple improvements over standard DiffSim such as partial backpropagation and graph mini-batching making DiffSim training stable and efficient. The potential of DiffSim is demonstrated in the successful discovery of transition paths for the Muller-Brown model potential as well as a benchmark chemical system - alanine dipeptide.
****

Preprocessors Matter! Realistic Decision-Based Attacks on Machine Learning Systems
Chawin Sitawarin, Florian Tramèr, Nicholas Carlini
Decision-based attacks construct adversarial examples against a machine learning (ML) model by making only hard-label queries. These attacks have mainly been applied directly to standalone neural networks. However, in practice, ML models are just one component of a larger learning system. We find that by adding a singl

e preprocessor in front of a classifier, state-of-the-art query-based attacks ar e up to seven× less effective at attacking a prediction pipeline than at attacki ng the model alone. We explain this discrepancy by the fact that most preprocess ors introduce some notion of invariance to the input space. Hence, attacks that are unaware of this invariance inevitably waste a large number of queries to re- discover or overcome it. We, therefore, develop techniques to (i) reverse-engine er the preprocessor and then (ii) use this extracted information to attack the e nd-to-end system. Our preprocessors extraction method requires only a few hundre d queries, and our preprocessor-aware attacks recover the same efficacy as when attacking the model alone. The code can be found at https://github.com/google-re search/preprocessor-aware-black-box-attack.

****

Invariance in Policy Optimisation and Partial Identifiability in Reward Learning

Joar Max Viktor Skalse, Matthew Farrugia-Roberts, Stuart Russell, Alessandro Aba te, Adam Gleave

It is often very challenging to manually design reward functions for complex, re al-world tasks. To solve this, one can instead use reward learning to infer a re ward function from data. However, there are often multiple reward functions that fit the data equally well, even in the infinite-data limit. This means that the reward function is only partially identifiable. In this work, we formally chara cterise the partial identifiability of the reward function given several popular reward learning data sources, including expert demonstrations and trajectory co mparisons. We also analyse the impact of this partial identifiability for severa l downstream tasks, such as policy optimisation. We unify our results in a frame work for comparing data sources and downstream tasks by their invariances, with implications for the design and selection of data sources for reward learning.

****

A Game-Theoretic Framework for Managing Risk in Multi-Agent Systems

Oliver Slumbers, David Henry Mguni, Stefano B Blumberg, Stephen Marcus Mcaleer, Yaodong Yang, Jun Wang

In order for agents in multi-agent systems (MAS) to be safe, they need to take i nto account the risks posed by the actions of other agents. However, the dominan t paradigm in game theory (GT) assumes that agents are not affected by risk from other agents and only strive to maximise their expected utility. For example, i n hybrid human-AI driving systems, it is necessary to limit large deviations in reward resulting from car crashes. Although there are equilibrium concepts in ga me theory that take into account risk aversion, they either assume that agents a re risk-neutral with respect to the uncertainty caused by the actions of other a gents, or they are not guaranteed to exist. We introduce a new GT-based Risk-Ave rse Equilibrium (RAE) that always produces a solution that minimises the potenti al variance in reward accounting for the strategy of other agents. Theoretically and empirically, we show RAE shares many properties with a Nash Equilibrium (NE), establishing convergence properties and generalising to risk-dominant NE in c ertain cases. To tackle large-scale problems, we extend RAE to the PSRO multi-ag ent reinforcement learning (MARL) framework. We empirically demonstrate the mini mum reward variance benefits of RAE in matrix games with high-risk outcomes. Res ults on MARL experiments show RAE generalises to risk-dominant NE in a trust dil emma game and that it reduces instances of crashing by 7x in an autonomous drivi ng setting versus the best performing baseline.

****

On the Effectiveness of Offline RL for Dialogue Response Generation

Paloma Sodhi, Felix Wu, Ethan R. Elenberg, Kilian Q Weinberger, Ryan Mcdonald

A common training technique for language models is teacher forcing (TF). TF atte mpts to match human language exactly, even though identical meanings can be expr essed in different ways. This motivates use of sequence-level objectives for dia logue response generation. In this paper, we study the efficacy of various offli ne reinforcement learning (RL) methods to maximize such objectives. We present a comprehensive evaluation across multiple datasets, models, and metrics. Offline RL shows a clear performance improvement over teacher forcing while not inducin g training instability or sacrificing practical training budgets.

****
Fair Densities via Boosting the Sufficient Statistics of Exponential Families
Alexander Soen, Hisham Husain, Richard Nock
We introduce a boosting algorithm to pre-process data for fairness. Starting fro m an initial fair but inaccurate distribution, our approach shifts towards bette r data fitting while still ensuring a minimal fairness guarantee. To do so, it l earns the sufficient statistics of an exponential family with boosting-compliant convergence. Importantly, we are able to theoretically prove that the learned d istribution will have a representation rate and statistical rate data fairness g uarantee. Unlike recent optimization based pre-processing methods, our approach can be easily adapted for continuous domain features. Furthermore, when the weak learners are specified to be decision trees, the sufficient statistics of the l earned distribution can be examined to provide clues on sources of (un)fairness. Empirical results are present to display the quality of result on real-world da ta.
****
The Dormant Neuron Phenomenon in Deep Reinforcement Learning
Ghada Sokar, Rishabh Agarwal, Pablo Samuel Castro, Utku Evci
In this work we identify the dormant neuron phenomenon in deep reinforcement lea rning, where an agent’s network suffers from an increasing number of inactive ne urons, thereby affecting network expressivity. We demonstrate the presence of th is phenomenon across a variety of algorithms and environments, and highlight its effect on learning. To address this issue, we propose a simple and effective me thod (ReDo) that Recycles Dormant neurons throughout training. Our experiments d emonstrate that ReDo maintains the expressive power of networks by reducing the number of dormant neurons and results in improved performance.
****
Abstracting Imperfect Information Away from Two-Player Zero-Sum Games
Samuel Sokota, Ryan D’Orazio, Chun Kai Ling, David J Wu, J Zico Kolter, Noam Bro wn
In their seminal work, Nayyar et al. (2013) showed that imperfect information ca n be abstracted away from common-payoff games by having players publicly announc e their policies as they play. This insight underpins sound solvers and decision -time planning algorithms for common-payoff games. Unfortunately, a naive applic ation of the same insight to two-player zero-sum games fails because Nash equili bria of the game with public policy announcements may not correspond to Nash equ ilibria of the original game. As a consequence, existing sound decision-time pla nning algorithms require complicated additional mechanisms that have unappealing properties. The main contribution of this work is showing that certain regulari zed equilibria do not possess the aforementioned non-correspondence problem—thus , computing them can be treated as perfect-information problems. Because these r egularized equilibria can be made arbitrarily close to Nash equilibria, our resu lt opens the door to a new perspective to solving two-player zero-sum games and yields a simplified framework for decision-time planning in two-player zero-sum games, void of the unappealing properties that plague existing decision-time pla nning approaches.
****
Meta-SAGE: Scale Meta-Learning Scheduled Adaptation with Guided Exploration for Mitigating Scale Shift on Combinatorial Optimization
Jiwoo Son, Minsu Kim, Hyeonah Kim, Jinkyoo Park
This paper proposes Meta-SAGE, a novel approach for improving the scalability of deep reinforcement learning models for combinatorial optimization (CO) tasks. O ur method adapts pre-trained models to larger-scale problems in test time by sug gesting two components: a scale meta-learner (SML) and scheduled adaptation with guided exploration (SAGE). First, SML transforms the context embedding for subs equent adaptation of SAGE based on scale information. Then, SAGE adjusts the mod el parameters dedicated to the context embedding for a specific instance. SAGE i ntroduces locality bias, which encourages selecting nearby locations to determin e the next location. The locality bias gradually decays as the model is adapted to the target instance. Results show that Meta-SAGE outperforms previous adaptat

ion methods and significantly improves scalability in representative CO tasks. Our source code is available at https://github.com/kaist-silab/meta-sage.
****

Consistency Models
Yang Song, Prafulla Dhariwal, Mark Chen, Ilya Sutskever
Diffusion models have significantly advanced the fields of image, audio, and video generation, but they depend on an iterative sampling process that causes slow generation. To overcome this limitation, we propose consistency models, a new family of models that generate high quality samples by directly mapping noise to data. They support fast one-step generation by design, while still allowing multistep sampling to trade compute for sample quality. They also support zero-shot data editing, such as image inpainting, colorization, and super-resolution, without requiring explicit training on these tasks. Consistency models can be trained either by distilling pre-trained diffusion models, or as standalone generative models altogether. Through extensive experiments, we demonstrate that they outperform existing distillation techniques for diffusion models in one- and few-step sampling, achieving the new state-of-the-art FID of 3.55 on CIFAR-10 and 6.20 on ImageNet 64x64 for one-step generation. When trained in isolation, consistency models become a new family of generative models that can outperform existing one-step, non-adversarial generative models on standard benchmarks such as CIFAR-10, ImageNet 64x64 and LSUN 256x256.
****

LipsNet: A Smooth and Robust Neural Network with Adaptive Lipschitz Constant for High Accuracy Optimal Control
Xujie Song, Jingliang Duan, Wenxuan Wang, Shengbo Eben Li, Chen Chen, Bo Cheng, Bo Zhang, Junqing Wei, Xiaoming Simon Wang
Deep reinforcement learning (RL) is a powerful approach for solving optimal control problems. However, RL-trained policies often suffer from the action fluctuation problem, where the consecutive actions significantly differ despite only slight state variations. This problem results in mechanical components' wear and tear and poses safety hazards. The action fluctuation is caused by the high Lipschitz constant of actor networks. To address this problem, we propose a neural network named LipsNet. We propose the Multi-dimensional Gradient Normalization (MGN) method, to constrain the Lipschitz constant of networks with multi-dimensional input and output. Benefiting from MGN, LipsNet achieves Lipschitz continuity, allowing smooth actions while preserving control performance by adjusting Lipschitz constant. LipsNet addresses the action fluctuation problem at network level rather than algorithm level, which can serve as actor networks in most RL algorithms, making it more flexible and user-friendly than previous works. Experiments demonstrate that LipsNet has good landscape smoothness and noise robustness, resulting in significantly smoother action compared to the Multilayer Perceptron.
****

Deep Perturbation Learning: Enhancing the Network Performance via Image Perturbations
Zifan Song, Xiao Gong, Guosheng Hu, Cairong Zhao
Image perturbation technique is widely used to generate adversarial examples to attack networks, greatly decreasing the performance of networks. Unlike the existing works, in this paper, we introduce a novel framework Deep Perturbation Learning (DPL), the new insights into understanding image perturbations, to enhance the performance of networks rather than decrease the performance. Specifically, we learn image perturbations to amend the data distribution of training set to improve the performance of networks. This optimization w.r.t data distribution is non-trivial. To approach this, we tactfully construct a differentiable optimization target w.r.t. image perturbations via minimizing the empirical risk. Then we propose an alternating optimization of the network weights and perturbations. DPL can easily be adapted to a wide spectrum of downstream tasks and backbone networks. Extensive experiments demonstrate the effectiveness of our DPL on 6 datasets (CIFAR-10, CIFAR100, ImageNet, MS-COCO, PASCAL VOC, and SBD) over 3 popular vision tasks (image classification, object detection, and semantic segmentation) with different backbone architectures (e.g., ResNet, MobileNet, and ViT).

****
Latent Traversals in Generative Models as Potential Flows
Yue Song, T. Anderson Keller, Nicu Sebe, Max Welling
Despite the significant recent progress in deep generative models, the underlying structure of their latent spaces is still poorly understood, thereby making the task of performing semantically meaningful latent traversals an open research challenge. Most prior work has aimed to solve this challenge by modeling latent structures linearly, and finding corresponding linear directions which result in ‘disentangled’ generations. In this work, we instead propose to model latent structures with a learned dynamic potential landscape, thereby performing latent traversals as the flow of samples down the landscape’s gradient. Inspired by physics, optimal transport, and neuroscience, these potential landscapes are learned as physically realistic partial differential equations, thereby allowing them to flexibly vary over both space and time. To achieve disentanglement, multiple potentials are learned simultaneously, and are constrained by a classifier to be distinct and semantically self-consistent. Experimentally, we demonstrate that our method achieves both more qualitatively and quantitatively disentangled trajectories than state-of-the-art baselines. Further, we demonstrate that our method can be integrated as a regularization term during training, thereby acting as an inductive bias towards the learning of structured representations, ultimately improving model likelihood on similarly structured data. Code is available at https://github.com/KingJamesSong/PDETraversal.
****
FedAvg Converges to Zero Training Loss Linearly for Overparameterized Multi-Layer Neural Networks
Bingqing Song, Prashant Khanduri, Xinwei Zhang, Jinfeng Yi, Mingyi Hong
Federated Learning (FL) is a distributed learning paradigm that allows multiple clients to learn a joint model by utilizing privately held data at each client. Significant research efforts have been devoted to develop advanced algorithms that deal with the situation where the data at individual clients have heterogeneous distributions. In this work, we show that data heterogeneity can be dealt from a different perspective. That is, by utilizing a certain overparameterized multi-layer neural network at each client, even the vanilla FedAvg (a.k.a. the Local SGD) algorithm can accurately optimize the training problem: When each client has a neural network with one wide layer of size N (where N is the number of total training samples), followed by layers of smaller widths, FedAvg converges linearly to a solution that achieves (almost) zero training loss, without requiring any assumptions on the clients’ data distributions. To our knowledge, this is the first work that demonstrates such resilience to data heterogeneity for FedAvg when trained on multi-layer neural networks. Our experiments also confirm that, neural networks of large size can achieve better and more stable performance for FL problems.
****
RGE: A Repulsive Graph Rectification for Node Classification via Influence
Jaeyun Song, Sungyub Kim, Eunho Yang
In real-world graphs, noisy connections are inevitable, which makes it difficult to obtain unbiased node representations. Among various attempts to resolve this problem, a method of estimating the counterfactual effects of these connectivities has recently attracted attention, which mainly uses influence functions for single graph elements (i.e., node and edge). However, in this paper, we argue that there is a strongly interacting group effect between the influences of graph elements due to their connectivity. In the same vein, we observe that edge groups connecting to the same train node exhibit significant differences in their influences, hence no matter how negative each is, removing them at once may have a rather negative effect as a group. Based on this motivation, we propose a new edge-removing strategy, Repulsive edge Group Elimination (RGE), that preferentially removes edges with no interference in groups. Empirically, we demonstrate that RGE consistently outperforms existing methods on the various benchmark datasets.
****

Importance Weighted Expectation-Maximization for Protein Sequence Design
Zhenqiao Song, Lei Li
Designing protein sequences with desired biological function is crucial in biology and chemistry. Recent machine learning methods use a surrogate sequence-function model to replace the expensive wet-lab validation. How can we efficiently generate diverse and novel protein sequences with high fitness? In this paper, we propose IsEM-Pro, an approach to generate protein sequences towards a given fitness criterion. At its core, IsEM-Pro is a latent generative model, augmented by combinatorial structure features from a separately learned Markov random fields (MRFs). We develop an Monte Carlo Expectation-Maximization method (MCEM) to learn the model. During inference, sampling from its latent space enhances diversity while its MRFs features guide the exploration in high fitness regions. Experiments on eight protein sequence design tasks show that our IsEM-Pro outperforms the previous best methods by at least 55% on average fitness score and generates more diverse and novel protein sequences.
****
Sketching for First Order Method: Efficient Algorithm for Low-Bandwidth Channel and Vulnerability
Zhao Song, Yitan Wang, Zheng Yu, Lichen Zhang
Sketching is one of the most fundamental tools in large-scale machine learning. It enables runtime and memory saving via randomly compressing the original large problem into lower dimensions. In this paper, we propose a novel sketching scheme for the first order method in large-scale distributed learning setting, such that the communication costs between distributed agents are saved while the convergence of the algorithms is still guaranteed. Given gradient information in a high dimension d, the agent passes the compressed information processed by a sketching matrix $R\in \mathbb{R}^{s\times d}$ with $s\ll d$, and the receiver de-compressed via the de-sketching matrix $R^\top$ to “recover” the information in original dimension. Using such a framework, we develop algorithms for federated learning with lower communication costs. However, such random sketching does not protect the privacy of local data directly. We show that the gradient leakage problem still exists after applying the sketching technique by presenting a specific gradient attack method. As a remedy, we prove rigorously that the algorithm will be differentially private by adding additional random noises in gradient information, which results in a both communication-efficient and differentially private first order approach for federated learning tasks. Our sketching scheme can be further generalized to other learning settings and might be of independent interest itself.
****
Sketching Meets Differential Privacy: Fast Algorithm for Dynamic Kronecker Projection Maintenance
Zhao Song, Xin Yang, Yuanyuan Yang, Lichen Zhang
Projection maintenance is one of the core data structure tasks. Efficient data structures for projection maintenance have led to recent breakthroughs in many convex programming algorithms. In this work, we further extend this framework to the Kronecker product structure. Given a constraint matrix ${\sf A}$ and a positive semi-definite matrix $W\in \mathbb{R}^{n\times n}$ with a sparse eigenbasis, we consider the task of maintaining the projection in the form of ${\sf B}^\top({\sf B}{\sf B}^\top)^{-1}{\sf B}$, where ${\sf B}={\sf A}(W\otimes I)$ or ${\sf B}={\sf A}(W^{1/2}\otimes W^{1/2})$. At each iteration, the weight matrix W receives a low rank change and we receive a new vector h. The goal is to maintain the projection matrix and answer the query ${\sf B}^\top({\sf B}{\sf B}^\top)^{-1}{\sf B}h$ with good approximation guarantees. We design a fast dynamic data structure for this task and it is robust against an adaptive adversary. Following the beautiful and pioneering work of [Beimel, Kaplan, Mansour, Nissim, Saranurak and Stemmer, STOC’22], we use tools from differential privacy to reduce the randomness required by the data structure and further improve the running time.
****
A Nearly-Optimal Bound for Fast Regression with $\ell_∞$ Guarantee
Zhao Song, Mingquan Ye, Junze Yin, Lichen Zhang

Given a matrix $A\in \mathbb{R}^{n\times d}$ and a vector $b\in \mathbb{R}^n$, we consider the regression problem with ℓ_∞ guarantees: finding a vector $x'\in \mathbb{R}^d$ such that $||x'-x^* ||_\infty \leq \frac{\epsilon}{\sqrt{d}}\cdot ||Ax^*-b||_2\cdot ||A^\dagger||$ with x^* being the optimal solution to the regression $||Ax-b||_2$. One popular approach for solving ℓ_2 regression problem is via sketching: picking a structured random matrix $S\in \mathbb{R}^{m\times n}$ with $m\ll n$ and SA can be quickly computed, solve the "sketched" regression problem $x'=\mathrm{argmin} ||SAx-Sb||_2$. In this paper, we show that in order to obtain such ℓ_∞ guarantee for ℓ_2 regression, one has to use sketching matrices that are dense. To the best of our knowledge, this is the first user case in which dense sketching matrices are necessary. On the algorithmic side, we prove that, there exists a distribution of dense sketching matrices with $m=\epsilon^{-2}d\log^3(n/\delta)$ such that solving the sketched regression problem gives the ℓ_∞ guarantee, with probability at least $1-\delta$. Moreover, the matrix SA can be computed in time $O(nd\log n)$. Our row count is nearly-optimal up to logarithmic factors, and significantly improves the result in [Price, Song and Woodruff, ICALP'17], in which $m=\Omega(\epsilon^{-2}d^{1+\gamma})$ for $\gamma\in (0, 1)$ is required. Moreover, we develop a novel analytical framework for ℓ_∞ guarantee regression that utilizes the Oblivious Coordinate-wise Embedding (OCE) property introduced in [Song and Yu, ICML'21]. Our analysis is much simpler and more general than that of [Price, Song and Woodruff, ICALP'17]. Leveraging this framework, we extend the ℓ_∞ guarantee regression result to dense sketching matrices for computing fast tensor product of vectors.

****

Loss-Guided Diffusion Models for Plug-and-Play Controllable Generation

Jiaming Song, Qinsheng Zhang, Hongxu Yin, Morteza Mardani, Ming-Yu Liu, Jan Kautz, Yongxin Chen, Arash Vahdat

We consider guiding denoising diffusion models with general differentiable loss functions in a plug-and-play fashion, enabling controllable generation without additional training. This paradigm, termed Loss-Guided Diffusion (LGD), can easily be integrated into all diffusion models and leverage various efficient samplers. Despite the benefits, the resulting guidance term is, unfortunately, an intractable integral and needs to be approximated. Existing methods compute the guidance term based on a point estimate. However, we show that such approaches have significant errors over the scale of the approximations. To address this issue, we propose a Monte Carlo method that uses multiple samples from a suitable distribution to reduce bias. Our method is effective in various synthetic and real-world settings, including image super-resolution, text or label-conditional image generation, and controllable motion synthesis. Notably, we show how our method can be applied to control a pretrained motion diffusion model to follow certain paths and avoid obstacles that are proven challenging to prior methods.

****

Differentiable Tree Operations Promote Compositional Generalization

Paul Soulos, Edward J Hu, Kate Mccurdy, Yunmo Chen, Roland Fernandez, Paul Smolensky, Jianfeng Gao

In the context of structure-to-structure transformation tasks, learning sequences of discrete symbolic operations poses significant challenges due to their non-differentiability. To facilitate the learning of these symbolic sequences, we introduce a differentiable tree interpreter that compiles high-level symbolic tree operations into subsymbolic matrix operations on tensors. We present a novel Differentiable Tree Machine (DTM) architecture that integrates our interpreter with an external memory and an agent that learns to sequentially select tree operations to execute the target transformation in an end-to-end manner. With respect to out-of-distribution compositional generalization on synthetic semantic parsing and language generation tasks, DTM achieves 100% while existing baselines such as Transformer, Tree Transformer, LSTM, and Tree2Tree LSTM achieve less than 30%. DTM remains highly interpretable in addition to its perfect performance.

****

Are labels informative in semi-supervised learning? Estimating and leveraging th

e missing-data mechanism.
Aude Sportisse, Hugo Schmutz, Olivier Humbert, Charles Bouveyron, Pierre-Alexandre Mattei
Semi-supervised learning is a powerful technique for leveraging unlabeled data to improve machine learning models, but it can be affected by the presence of “informative" labels, which occur when some classes are more likely to be labeled than others. In the missing data literature, such labels are called missing not at random. In this paper, we propose a novel approach to address this issue by estimating the missing-data mechanism and using inverse propensity weighting to debias any SSL algorithm, including those using data augmentation. We also propose a likelihood ratio test to assess whether or not labels are indeed informative. Finally, we demonstrate the performance of the proposed methods on different datasets, in particular on two medical datasets for which we design pseudo-realistic missing data scenarios.
****
Linear Causal Disentanglement via Interventions
Chandler Squires, Anna Seigal, Salil S Bhate, Caroline Uhler
Causal disentanglement seeks a representation of data involving latent variables that are related via a causal model. A representation is identifiable if both the latent model and the transformation from latent to observed variables are unique. In this paper, we study observed variables that are a linear transformation of a linear latent causal model. Data from interventions are necessary for identifiability: if one latent variable is missing an intervention, we show that there exist distinct models that cannot be distinguished. Conversely, we show that a single intervention on each latent variable is sufficient for identifiability. Our proof uses a generalization of the RQ decomposition of a matrix that replaces the usual orthogonal and upper triangular conditions with analogues depending on a partial order on the rows of the matrix, with partial order determined by a latent causal model. We corroborate our theoretical results with a method for causal disentanglement. We show that the method accurately recovers a latent causal model on synthetic and semi-synthetic data and we illustrate a use case on a dataset of single-cell RNA sequencing measurements.
****
Generating Language Corrections for Teaching Physical Control Tasks
Megha Srivastava, Noah Goodman, Dorsa Sadigh
AI assistance continues to help advance applications in education, from language learning to intelligent tutoring systems, yet current methods for providing students feedback are still quite limited. Most automatic feedback systems either provide binary correctness feedback, which may not help a student understand how to improve, or require hand-coding feedback templates, which may not generalize to new domains. This can be particularly challenging for physical control tasks, where the rich diversity in student behavior and specialized domains make it challenging to leverage general-purpose assistive tools for providing feedback. We design and build CORGI, a model trained to generate language corrections for physical control tasks, such as learning to ride a bike. CORGI takes in as input a pair of student and expert trajectories, and then generates natural language corrections to help the student improve. We collect and train CORGI over data from three diverse physical control tasks (drawing, steering, and joint movement). Through both automatic and human evaluations, we show that CORGI can (i) generate valid feedback for novel student trajectories, (ii) outperform baselines on domains with novel control dynamics, and (iii) improve student learning in an interactive drawing task.
****
FaDIn: Fast Discretized Inference for Hawkes Processes with General Parametric Kernels
Guillaume Staerman, Cédric Allain, Alexandre Gramfort, Thomas Moreau
Temporal point processes (TPP) are a natural tool for modeling event-based data. Among all TPP models, Hawkes processes have proven to be the most widely used, mainly due to their adequate modeling for various applications, particularly when considering exponential or non-parametric kernels. Although non-parametric ker

nels are an option, such models require large datasets. While exponential kernel
s are more data efficient and relevant for specific applications where events im
mediately trigger more events, they are ill-suited for applications where latenc
ies need to be estimated, such as in neuroscience. This work aims to offer an ef
ficient solution to TPP inference using general parametric kernels with finite s
upport. The developed solution consists of a fast ℓ_2 gradient-based solver
 leveraging a discretized version of the events. After theoretically supporting
the use of discretization, the statistical and computational efficiency of the n
ovel approach is demonstrated through various numerical experiments. Finally, th
e method’s effectiveness is evaluated by modeling the occurrence of stimuli-indu
ced patterns from brain signals recorded with magnetoencephalography (MEG). Give
n the use of general parametric kernels, results show that the proposed approach
 leads to an improved estimation of pattern latency than the state-of-the-art.
****

Partial Optimality in Cubic Correlation Clustering
David Stein, Silvia Di Gregorio, Bjoern Andres
The higher-order correlation clustering problem is an expressive model, and rece
ntly, local search heuristics have been proposed for several applications. Certi
fying optimality, however, is NP-hard and practically hampered already by the co
mplexity of the problem statement. Here, we focus on establishing partial optima
lity conditions for the special case of complete graphs and cubic objective func
tions. In addition, we define and implement algorithms for testing these conditi
ons and examine their effect numerically, on two datasets.
****

MODeL: Memory Optimizations for Deep Learning
Benoit Steiner, Mostafa Elhoushi, Jacob Kahn, James Hegarty
The size of deep neural networks has grown exponentially in recent years. Unfort
unately, hardware devices have not kept pace with the rapidly increasing memory
 requirements. To cope with this, researchers have proposed various techniques in
cluding spilling, rematerialization, reduced precision training, model pruning,
and so on. However, these approaches suffer from various limitations, such as in
creasing training time, affecting model accuracy, or requiring extensive manual
 modifications to the neural networks. We present MODeL, an algorithm that optimi
zes the lifetime and memory location of the tensors used to train neural network
s. Our method automatically reduces the memory usage of existing neural networks
 without any of the drawbacks of other techniques. We formulate the problem as a
 joint integer linear program (ILP). We present several techniques to simplify t
he encoding of the problem, and enable our approach to scale to the size of stat
e-of-the-art neural networks using an off-the-shelf ILP solver. We experimentall
y demonstrate that MODeL only takes seconds to allow the training of neural netw
orks using 30% less memory on average.
****

Improving Expert Predictions with Conformal Prediction
Eleni Straitouri, Lequn Wang, Nastaran Okati, Manuel Gomez Rodriguez
Automated decision support systems promise to help human experts solve multiclas
s classification tasks more efficiently and accurately. However, existing system
s typically require experts to understand when to cede agency to the system or w
hen to exercise their own agency. Otherwise, the experts may be better off solvi
ng the classification tasks on their own. In this work, we develop an automated
 decision support system that, by design, does not require experts to understand
 when to trust the system to improve performance. Rather than providing (single)
 label predictions and letting experts decide when to trust these predictions, ou
r system provides sets of label predictions constructed using conformal predicti
on—prediction sets—and forcefully asks experts to predict labels from these sets
. By using conformal prediction, our system can precisely trade-off the probabil
ity that the true label is not in the prediction set, which determines how frequ
ently our system will mislead the experts, and the size of the prediction set, w
hich determines the difficulty of the classification task the experts need to so
lve using our system. In addition, we develop an efficient and near-optimal sear
ch method to find the conformal predictor under which the experts benefit the mo

st from using our system. Simulation experiments using synthetic and real expert predictions demonstrate that our system may help experts make more accurate predictions and is robust to the accuracy of the classifier the conformal predictor relies on.
****
Lookahead When It Matters: Adaptive Non-causal Transformers for Streaming Neural Transducers
Grant Strimel, Yi Xie, Brian John King, Martin Radfar, Ariya Rastrow, Athanasios Mouchtaris
Streaming speech recognition architectures are employed for low-latency, real-time applications. Such architectures are often characterized by their causality. Causal architectures emit tokens at each frame, relying only on current and past signal, while non-causal models are exposed to a window of future frames at each step to increase predictive accuracy. This dichotomy amounts to a trade-off for real-time Automatic Speech Recognition (ASR) system design: profit from the low-latency benefit of strictly-causal architectures while accepting predictive performance limitations, or realize the modeling benefits of future-context models accompanied by their higher latency penalty. In this work, we relax the constraints of this choice and present the Adaptive Non-Causal Attention Transducer (ANCAT). Our architecture is non-causal in the traditional sense, but executes in a low-latency, streaming manner by dynamically choosing when to rely on future context and to what degree within the audio stream. The resulting mechanism, when coupled with our novel regularization algorithms, delivers comparable accuracy to non-causal configurations while improving significantly upon latency, closing the gap with their causal counterparts. We showcase our design experimentally by reporting comparative ASR task results with measures of accuracy and latency on both publicly accessible and production-scale, voice-assistant datasets.
****
Kernel QuantTree
Diego Stucchi, Paolo Rizzo, Nicolò Folloni, Giacomo Boracchi
We present Kernel QuantTree (KQT), a non-parametric change detection algorithm that monitors multivariate data through a histogram. KQT constructs a nonlinear partition of the input space that matches pre-defined target probabilities and specifically promotes compact bins adhering to the data distribution, resulting in a powerful detection algorithm. We prove two key theoretical advantages of KQT: i) statistics defined over the KQT histogram do not depend on the stationary data distribution ϕ_0, so detection thresholds can be set a priori to control false positive rate, and ii) thanks to the kernel functions adopted, the KQT monitoring scheme is invariant to the roto-translation of the input data. Consequently, KQT does not require any preprocessing step like PCA. Our experiments show that KQT achieves superior detection power than non-parametric state-of-the-art change detection methods, and can reliably control the false positive rate.
****
Topologically Faithful Image Segmentation via Induced Matching of Persistence Barcodes
Nico Stucki, Johannes C. Paetzold, Suprosanna Shit, Bjoern Menze, Ulrich Bauer
Segmentation models predominantly optimize pixel-overlap-based loss, an objective that is actually inadequate for many segmentation tasks. In recent years, their limitations fueled a growing interest in topology-aware methods, which aim to recover the topology of the segmented structures. However, so far, existing methods only consider global topological properties, ignoring the need to preserve topological features spatially, which is crucial for accurate segmentation. We introduce the concept of induced matchings from persistent homology to achieve a spatially correct matching between persistence barcodes in a segmentation setting. Based on this concept, we define the Betti matching error as an interpretable, topologically and feature-wise accurate metric for image segmentations, which resolves the limitations of the Betti number error. Our Betti matching error is differentiable and efficient to use as a loss function. We demonstrate that it improves the topological performance of segmentation networks significantly across six diverse datasets while preserving the performance with respect to tradition

al scores. Our code is publicly available (https://github.com/nstucki/Betti-matching/).
****
Towards Robust Graph Incremental Learning on Evolving Graphs
Junwei Su, Difan Zou, Zijun Zhang, Chuan Wu
Incremental learning is a machine learning approach that involves training a model on a sequence of tasks, rather than all tasks at once. This ability to learn incrementally from a stream of tasks is crucial for many real-world applications. However, incremental learning is a challenging problem on graph-structured data, as many graph-related problems involve prediction tasks for each individual node, known as Node-wise Graph Incremental Learning (NGIL). This introduces non-independent and non-identically distributed characteristics in the sample data generation process, making it difficult to maintain the performance of the model as new tasks are added. In this paper, we focus on the inductive NGIL problem, which accounts for the evolution of graph structure (structural shift) induced by emerging tasks. We provide a formal formulation and analysis of the problem, and propose a novel regularization-based technique called Structural-Shift-Risk-Mitigation (SSRM) to mitigate the impact of the structural shift on catastrophic forgetting of the inductive NGIL problem. We show that the structural shift can lead to a shift in the input distribution for the existing tasks, and further lead to an increased risk of catastrophic forgetting. Through comprehensive empirical studies with several benchmark datasets, we demonstrate that our proposed method, Structural-Shift-Risk-Mitigation (SSRM), is flexible and easy to adapt to improve the performance of state-of-the-art GNN incremental learning frameworks in the inductive setting.
****
DUET: 2D Structured and Approximately Equivariant Representations
Xavier Suau, Federico Danieli, T. Anderson Keller, Arno Blaas, Chen Huang, Jason Ramapuram, Dan Busbridge, Luca Zappella
Multiview Self-Supervised Learning (MSSL) is based on learning invariances with respect to a set of input transformations. However, invariance partially or totally removes transformation-related information from the representations, which might harm performance for specific downstream tasks that require such information. We propose 2D strUctured and EquivarianT representations (coined DUET), which are 2d representations organized in a matrix structure, and equivariant with respect to transformations acting on the input data. DUET representations maintain information about an input transformation, while remaining semantically expressive. Compared to SimCLR (Chen et al., 2020) (unstructured and invariant) and ESSL (Dangovski et al., 2022) (unstructured and equivariant), the structured and equivariant nature of DUET representations enables controlled generation with lower reconstruction error, while controllability is not possible with SimCLR or ESSL. DUET also achieves higher accuracy for several discriminative tasks, and improves transfer learning.
****
Long-Tailed Recognition by Mutual Information Maximization between Latent Features and Ground-Truth Labels
Min-Kook Suh, Seung-Woo Seo
Although contrastive learning methods have shown prevailing performance on a variety of representation learning tasks, they encounter difficulty when the training dataset is long-tailed. Many researchers have combined contrastive learning and a logit adjustment technique to address this problem, but the combinations are done ad-hoc and a theoretical background has not yet been provided. The goal of this paper is to provide the background and further improve the performance. First, we show that the fundamental reason contrastive learning methods struggle with long-tailed tasks is that they try to maximize the mutual information between latent features and input data. As ground-truth labels are not considered in the maximization, they are not able to address imbalances between classes. Rather, we interpret the long-tailed recognition task as a mutual information maximization between latent features and ground-truth labels. This approach integrates contrastive learning and logit adjustment seamlessly to derive a loss function t

hat shows state-of-the-art performance on long-tailed recognition benchmarks. It also demonstrates its efficacy in image segmentation tasks, verifying its versatility beyond image classification. Code is available at https://github.com/bluecdm/Long-tailed-recognition.

****

Adversarial Learning of Distributional Reinforcement Learning

Yang Sui, Yukun Huang, Hongtu Zhu, Fan Zhou

Reinforcement learning (RL) has made significant advancements in artificial intelligence. However, its real-world applications are limited due to differences between simulated environments and the actual world. Consequently, it is crucial to systematically analyze how each component of the RL system can affect the final model performance. In this study, we propose an adversarial learning framework for distributional reinforcement learning, which adopts the concept of influence measure from the statistics community. This framework enables us to detect performance loss caused by either the internal policy structure or the external state observation. The proposed influence measure is based on information geometry and has desirable properties of invariance. We demonstrate that the influence measure is useful for three diagnostic tasks: identifying fragile states in trajectories, determining the instability of the policy architecture, and pinpointing anomalously sensitive policy parameters.

****

Distilling Internet-Scale Vision-Language Models into Embodied Agents

Theodore Sumers, Kenneth Marino, Arun Ahuja, Rob Fergus, Ishita Dasgupta

Instruction-following agents must ground language into their observation and action spaces. Learning to ground language is challenging, typically requiring domain-specific engineering or large quantities of human interaction data. To address this challenge, we propose using pretrained vision-language models (VLMs) to supervise embodied agents. We combine ideas from model distillation and hindsight experience replay (HER), using a VLM to retroactively generate language describing the agent’s behavior. Simple prompting allows us to control the supervision signal, teaching an agent to interact with novel objects based on their names (e.g., planes) or their features (e.g., colors) in a 3D rendered environment. Fewshot prompting lets us teach abstract category membership, including pre-existing categories (food vs toys) and ad-hoc ones (arbitrary preferences over objects). Our work outlines a new and effective way to use internet-scale VLMs, repurposing the generic language grounding acquired by such models to teach task-relevant groundings to embodied agents.

****

Vector-Valued Control Variates

Zhuo Sun, Alessandro Barp, Francois-Xavier Briol

Control variates are variance reduction tools for Monte Carlo estimators. They can provide significant variance reduction, but usually require a large number of samples, which can be prohibitive when sampling or evaluating the integrand is computationally expensive. Furthermore, there are many scenarios where we need to compute multiple related integrals simultaneously or sequentially, which can further exacerbate computational costs. In this paper, we propose vector-valued control variates, an extension of control variates which can be used to reduce the variance of multiple Monte Carlo estimators jointly. This allows for the transfer of information across integration tasks, and hence reduces the need for a large number of samples. We focus on control variates based on kernel interpolants and our novel construction is obtained through a generalised Stein identity and the development of novel matrix-valued Stein reproducing kernels. We demonstrate our methodology on a range of problems including multifidelity modelling, Bayesian inference for dynamical systems, and model evidence computation through thermodynamic integration.

****

MetaModulation: Learning Variational Feature Hierarchies for Few-Shot Learning with Fewer Tasks

Wenfang Sun, Yingjun Du, Xiantong Zhen, Fan Wang, Ling Wang, Cees G. M. Snoek

Meta-learning algorithms are able to learn a new task using previously learned k

nowledge, but they often require a large number of meta-training tasks which may not be readily available. To address this issue, we propose a method for few-shot learning with fewer tasks, which we call MetaModulation. The key idea is to use a neural network to increase the density of the meta-training tasks by modulating batch normalization parameters during meta-training. Additionally, we modify parameters at various neural network levels, rather than just a single layer, to increase task diversity. To account for the uncertainty caused by the reduced number of training tasks, we propose a variational MetaModulation where the modulation parameters are treated as latent variables. We also introduce learning variational feature hierarchies by the variational MetaModulation, which modulates features at all layers and can take into account task uncertainty and generate more diverse tasks. The ablation studies illustrate the advantages of utilizing a learnable task modulation at different levels and demonstrate the benefit of incorporating probabilistic variants in few-task meta-learning. Our MetaModulation and its variational variants consistently outperform state-of-the-art alternatives on four few-task meta-learning benchmarks.

****

Revisiting Sampling for Combinatorial Optimization

Haoran Sun, Katayoon Goshvadi, Azade Nova, Dale Schuurmans, Hanjun Dai

Sampling approaches like Markov chain Monte Carlo were once popular for combinatorial optimization, but the inefficiency of classical methods and the need for problem-specific designs curtailed ongoing development. Recent work has favored data-driven approaches that mitigate the need for hand-craft heuristics, but these are often not usable as out-of-the-box solvers due to dependence on in-distribution training and limited scalability to large instances. In this paper, we revisit the idea of using sampling for combinatorial optimization, motivated by the significant recent advances of gradient-based discrete MCMC and new techniques for parallel neighborhood exploration on accelerators. Remarkably, we find that modern sampling strategies can leverage landscape information to provide general-purpose solvers that require no training and yet are competitive with state of the art combinatorial solvers. In particular, experiments on cover vertex selection, graph partition and routing demonstrate better speed-quality trade-offs over current learning based approaches, and sometimes even superior performance to commercial solvers and specialized algorithms.

****

What Makes Entities Similar? A Similarity Flooding Perspective for Multi-sourced Knowledge Graph Embeddings

Zequn Sun, Jiacheng Huang, Xiaozhou Xu, Qijin Chen, Weijun Ren, Wei Hu

Joint representation learning over multi-sourced knowledge graphs (KGs) yields transferable and expressive embeddings that improve downstream tasks. Entity alignment (EA) is a critical step in this process. Despite recent considerable research progress in embedding-based EA, how it works remains to be explored. In this paper, we provide a similarity flooding perspective to explain existing translation-based and aggregation-based EA models. We prove that the embedding learning process of these models actually seeks a fixpoint of pairwise similarities between entities. We also provide experimental evidence to support our theoretical analysis. We propose two simple but effective methods inspired by the fixpoint computation in similarity flooding, and demonstrate their effectiveness on benchmark datasets. Our work bridges the gap between recent embedding-based models and the conventional similarity flooding algorithm. It would improve our understanding of and increase our faith in embedding-based EA.

****

Maximum Optimality Margin: A Unified Approach for Contextual Linear Programming and Inverse Linear Programming

Chunlin Sun, Shang Liu, Xiaocheng Li

In this paper, we study the predict-then-optimize problem where the output of a machine learning prediction task is used as the input of some downstream optimization problem, say, the objective coefficient vector of a linear program. The problem is also known as predictive analytics or contextual linear programming. The existing approaches largely suffer from either (i) optimization intractability

(a non-convex objective function)/statistical inefficiency (a suboptimal generalization bound) or (ii) requiring strong condition(s) such as no constraint or loss calibration. We develop a new approach to the problem called maximum optimality margin which designs the machine learning loss function by the optimality condition of the downstream optimization. The max-margin formulation enjoys both computational efficiency and good theoretical properties for the learning procedure. More importantly, our new approach only needs the observations of the optimal solution in the training data rather than the objective function, which makes it a new and natural approach to the inverse linear programming problem under both contextual and context-free settings; we also analyze the proposed method under both offline and online settings, and demonstrate its performance using numerical experiments.

****

Tensor Gaussian Process with Contraction for Multi-Channel Imaging Analysis

Hu Sun, Ward Manchester, Meng Jin, Yang Liu, Yang Chen

Multi-channel imaging data is a prevalent data format in scientific fields such as astronomy and biology. The structured information and the high dimensionality of these 3-D tensor data makes the analysis an intriguing but challenging topic for statisticians and practitioners. The low-rank scalar-on-tensor regression model, in particular, has received widespread attention and has been re-formulated as a tensor Gaussian Process (Tensor-GP) model with multi-linear kernel in Yu et al. (2018). In this paper, we extend the Tensor-GP model by introducing an integrative dimensionality reduction technique, called tensor contraction, with a Tensor-GP for a scalar-on-tensor regression task with multi-channel imaging data. This is motivated by the solar flare forecasting problem with high dimensional multi-channel imaging data. We first estimate a latent, reduced-size tensor for each data tensor and then apply a multi-linear Tensor-GP on the latent tensor data for prediction. We introduce an anisotropic total-variation regularization when conducting the tensor contraction to obtain a sparse and smooth latent tensor. We then propose an alternating proximal gradient descent algorithm for estimation. We validate our approach via extensive simulation studies and applying it to the solar flare forecasting problem.

****

MABe22: A Multi-Species Multi-Task Benchmark for Learned Representations of Behavior

Jennifer J. Sun, Markus Marks, Andrew Wesley Ulmer, Dipam Chakraborty, Brian Geuther, Edward Hayes, Heng Jia, Vivek Kumar, Sebastian Oleszko, Zachary Partridge, Milan Peelman, Alice Robie, Catherine E Schretter, Keith Sheppard, Chao Sun, Param Uttarwar, Julian Morgan Wagner, Erik Werner, Joseph Parker, Pietro Perona, Yisong Yue, Kristin Branson, Ann Kennedy

We introduce MABe22, a large-scale, multi-agent video and trajectory benchmark to assess the quality of learned behavior representations. This dataset is collected from a variety of biology experiments, and includes triplets of interacting mice (4.7 million frames video+pose tracking data, 10 million frames pose only), symbiotic beetle-ant interactions (10 million frames video data), and groups of interacting flies (4.4 million frames of pose tracking data). Accompanying these data, we introduce a panel of real-life downstream analysis tasks to assess the quality of learned representations by evaluating how well they preserve information about the experimental conditions (e.g. strain, time of day, optogenetic stimulation) and animal behavior. We test multiple state-of-the-art self-supervised video and trajectory representation learning methods to demonstrate the use of our benchmark, revealing that methods developed using human action datasets do not fully translate to animal datasets. We hope that our benchmark and dataset encourage a broader exploration of behavior representation learning methods across species and settings.

****

Dynamic Regularized Sharpness Aware Minimization in Federated Learning: Approaching Global Consistency and Smooth Landscape

Yan Sun, Li Shen, Shixiang Chen, Liang Ding, Dacheng Tao

In federated learning (FL), a cluster of local clients are chaired under the coo

rdination of the global server and cooperatively train one model with privacy pr
otection. Due to the multiple local updates and the isolated non-iid dataset, cl
ients are prone to overfit into their own optima, which extremely deviates from
the global objective and significantly undermines the performance. Most previous
 works only focus on enhancing the consistency between the local and global obje
ctives to alleviate this prejudicial client drifts from the perspective of the o
ptimization view, whose performance would be prominently deteriorated on the hig
h heterogeneity. In this work, we propose a novel and general algorithm FedSMOO
by jointly considering the optimization and generalization targets to efficientl
y improve the performance in FL. Concretely, FedSMOO adopts a dynamic regularize
r to guarantee the local optima towards the global objective, which is meanwhile
 revised by the global Sharpness Aware Minimization (SAM) optimizer to search fo
r the consistent flat minima. Our theoretical analysis indicates that FedSMOO ac
hieves fast $\mathcal{O}(1/T)$ convergence rate with low generalization bound. E
xtensive numerical studies are conducted on the real-world dataset to verify its
 peerless efficiency and excellent generality.
****

When and How Does Known Class Help Discover Unknown Ones? Provable Understanding
 Through Spectral Analysis
Yiyou Sun, Zhenmei Shi, Yingyu Liang, Yixuan Li
Novel Class Discovery (NCD) aims at inferring novel classes in an unlabeled set
by leveraging prior knowledge from a labeled set with known classes. Despite its
 importance, there is a lack of theoretical foundations for NCD. This paper brid
ges the gap by providing an analytical framework to formalize and investigate wh
en and how known classes can help discover novel classes. Tailored to the NCD pr
oblem, we introduce a graph-theoretic representation that can be learned by a no
vel NCD Spectral Contrastive Loss (NSCL). Minimizing this objective is equivalen
t to factorizing the graph’s adjacency matrix, which allows us to derive a prova
ble error bound and provide the sufficient and necessary condition for NCD. Empi
rically, NSCL can match or outperform several strong baselines on common benchma
rk datasets, which is appealing for practical usage while enjoying theoretical g
uarantees.
****

Learning Prescriptive ReLU Networks
Wei Sun, Asterios Tsiourvas
We study the problem of learning optimal policy from a set of discrete treatment
 options using observational data. We propose a piecewise linear neural network
model that can balance strong prescriptive performance and interpretability, whi
ch we refer to as the prescriptive ReLU network, or P-ReLU. We show analytically
 that this model (i) partitions the input space into disjoint polyhedra, where a
ll instances that belong to the same partition receive the same treatment, and (
ii) can be converted into an equivalent prescriptive tree with hyperplane splits
 for interpretability. We demonstrate the flexibility of the P-ReLU network as c
onstraints can be easily incorporated with minor modifications to the architectu
re. Through experiments, we validate the superior prescriptive accuracy of P-ReL
U against competing benchmarks. Lastly, we present examples of prescriptive tree
s extracted from trained P-ReLUs using a real-world dataset, for both the uncons
trained and constrained scenarios.
****

All in a Row: Compressed Convolution Networks for Graphs
Junshu Sun, Shuhui Wang, Xinzhe Han, Zhe Xue, Qingming Huang
Compared to Euclidean convolution, existing graph convolution methods generally
fail to learn diverse convolution operators under limited parameter scales and d
epend on additional treatments of multi-scale feature extraction. The challenges
 of generalizing Euclidean convolution to graphs arise from the irregular struct
ure of graphs. To bridge the gap between Euclidean space and graph space, we pro
pose a differentiable method for regularization on graphs that applies permutati
ons to the input graphs. The permutations constrain all nodes in a row regardles
s of their input order and therefore enable the flexible generalization of Eucli
dean convolution. Based on the regularization of graphs, we propose Compressed C

onvolution Network (CoCN) for hierarchical graph representation learning. CoCN f
ollows the local feature learning and global parameter sharing mechanisms of Con
volution Neural Networks. The whole model can be trained end-to-end and is able
to learn both individual node features and the corresponding structure features.
 We validate CoCN on several node classification and graph classification benchm
arks. CoCN achieves superior performance over competitive convolutional GNNs and
 graph pooling models. Codes are available at https://github.com/sunjss/CoCN.
****

Momentum Ensures Convergence of SIGNSGD under Weaker Assumptions

Tao Sun, Qingsong Wang, Dongsheng Li, Bao Wang

Sign Stochastic Gradient Descent (signSGD) is a communication-efficient stochast
ic algorithm that only uses the sign information of the stochastic gradient to u
pdate the model’s weights. However, the existing convergence theory of signSGD e
ither requires increasing batch sizes during training or assumes the gradient no
ise is symmetric and unimodal. Error feedback has been used to guarantee the con
vergence of signSGD under weaker assumptions at the cost of communication overhe
ad. This paper revisits the convergence of signSGD and proves that momentum can
remedy signSGD under weaker assumptions than previous techniques; in particular,
 our convergence theory does not require the assumption of bounded stochastic gr
adient or increased batch size. Our results resonate with echoes of previous emp
irical results where, unlike signSGD, signSGD with momentum maintains good perfo
rmance even with small batch sizes. Another new result is that signSGD with mome
ntum can achieve an improved convergence rate when the objective function is sec
ond-order smooth. We further extend our theory to signSGD with major vote and fe
derated learning.
****

A Critical Revisit of Adversarial Robustness in 3D Point Cloud Recognition with
Diffusion-Driven Purification

Jiachen Sun, Jiongxiao Wang, Weili Nie, Zhiding Yu, Zhuoqing Mao, Chaowei Xiao

3D point clouds serve as a crucial data representation in numerous real-world ap
plications such as autonomous driving, robotics, and medical imaging. While the
advancements in deep learning have spurred the utilization of 3D point clouds, d
eep models are notoriously vulnerable to adversarial attacks. Various defense so
lutions have been proposed to build robust models against adversarial attacks. I
n this work, we pinpoint a major limitation of the leading empirical defense, ad
versarial training, when applied to 3D point cloud models: gradient obfuscation,
 which significantly hampers robustness against potent attacks. To bridge the ga
p, we propose PointDP, a purification strategy that leverages diffusion models t
o defend against 3D adversarial attacks. Since PointDP does not rely on predefin
ed adversarial examples for training, it can defend against a variety of threats
. We conduct a comprehensive evaluation of PointDP across six representative 3D
point cloud architectures, employing sixteen strong and adaptive attacks to mani
fest its foundational robustness. Our evaluation shows that PointDP achieves sig
nificantly better (i.e., 12.6%-40.3%) adversarial robustness than state-of-the-a
rt methods under strong attacks bounded by different ℓ_p norms.
****

SDDM: Score-Decomposed Diffusion Models on Manifolds for Unpaired Image-to-Image
 Translation

Shikun Sun, Longhui Wei, Junliang Xing, Jia Jia, Qi Tian

Recent score-based diffusion models (SBDMs) show promising results in unpaired i
mage-to-image translation (I2I). However, existing methods, either energy-based
or statistically-based, provide no explicit form of the interfered intermediate
generative distributions. This work presents a new score-decomposed diffusion mo
del (SDDM) on manifolds to explicitly optimize the tangled distributions during
image generation. SDDM derives manifolds to make the distributions of adjacent t
ime steps separable and decompose the score function or energy guidance into an
image "denoising" part and a content "refinement" part. To refine the image in t
he same noise level, we equalize the refinement parts of the score function and
energy guidance, which permits multi-objective optimization on the manifold. We
also leverage the block adaptive instance normalization module to construct mani

folds with lower dimensions but still concentrated with the perturbed reference image. SDDM outperforms existing SBDM-based methods with much fewer diffusion steps on several I2I benchmarks.
****
A Neural PDE Solver with Temporal Stencil Modeling
Zhiqing Sun, Yiming Yang, Shinjae Yoo
Numerical simulation of non-linear partial differential equations plays a crucial role in modeling physical science and engineering phenomena, such as weather, climate, and aerodynamics. Recent Machine Learning (ML) models trained on low-resolution spatio-temporal signals have shown new promises in capturing important dynamics in high-resolution signals, under the condition that the models can effectively recover the missing details. However, this study shows that significant information is often lost in the low-resolution down-sampled features. To address such issues, we propose a new approach, namely Temporal Stencil Modeling (TSM), which combines the strengths of advanced time-series sequence modeling (with the HiPPO features) and state-of-the-art neural PDE solvers (with learnable stencil modeling). TSM aims to recover the lost information from the PDE trajectories and can be regarded as a temporal generalization of classic finite volume methods such as WENO. Our experimental results show that TSM achieves the new state-of-the-art simulation accuracy for 2-D incompressible Navier-Stokes turbulent flows: it significantly outperforms the previously reported best results by 19.9% in terms of the highly-correlated duration time, and reduces the inference latency into 80%. We also show a strong generalization ability of the proposed method to various out-of-distribution turbulent flow settings, as well as lower resolution or 1-D / 3-D settings. Our code is available at https://github.com/Edward-Sun/TSM-PDE .
****
Feature Expansion for Graph Neural Networks
Jiaqi Sun, Lin Zhang, Guangyi Chen, Peng Xu, Kun Zhang, Yujiu Yang
Graph neural networks aim to learn representations for graph-structured data and show impressive performance in node classification. Recently, many methods have studied the representations of GNNs from the perspective of optimization goals and spectral graph theory. However, the feature space that dominates representation learning has not been systematically studied in graph neural networks. In this paper, we propose to fill this gap by analyzing the feature space of both spatial and spectral models. We decompose graph neural networks into determined feature spaces and trainable weights, providing the convenience of studying the feature space explicitly using matrix space analysis. In particular, we find theoretically that the feature space tends to be linearly correlated due to repeated aggregations. In this case, the feature space is bounded by the poor representation of shared weights or the limited dimensionality of node attributes in existing models, leading to poor performance. Motivated by these findings, we propose 1) feature subspaces flattening and 2) structural principal components to expand the feature space. Extensive experiments verify the effectiveness of our proposed more comprehensive feature space, with comparable inference time to the baseline, and demonstrate its efficient convergence capability.
****
Model-Bellman Inconsistency for Model-based Offline Reinforcement Learning
Yihao Sun, Jiaji Zhang, Chengxing Jia, Haoxin Lin, Junyin Ye, Yang Yu
For offline reinforcement learning (RL), model-based methods are expected to be data-efficient as they incorporate dynamics models to generate more data. However, due to inevitable model errors, straightforwardly learning a policy in the model typically fails in the offline setting. Previous studies have incorporated conservatism to prevent out-of-distribution exploration. For example, MOPO penalizes rewards through uncertainty measures from predicting the next states, which we have discovered are loose bounds of the ideal uncertainty, i.e., the Bellman error. In this work, we propose MOdel-Bellman Inconsistency penalized offLinE Policy Optimization (MOBILE), a novel uncertainty-driven offline RL algorithm. MOBILE conducts uncertainty quantification through the inconsistency of Bellman estimations under an ensemble of learned dynamics models, which can be a better app

roximator to the true Bellman error, and penalizes the Bellman estimation based on this uncertainty. Empirically we have verified that our proposed uncertainty quantification can be significantly closer to the true Bellman error than the co mpared methods. Consequently, MOBILE outperforms prior offline RL approaches on most tasks of D4RL and NeoRL benchmarks.
****

Inflow, Outflow, and Reciprocity in Machine Learning
Mukund Sundararajan, Walid Krichene
Data is pooled across entities (individuals or enterprises) to create machine le arning models, and sometimes, the entities that contribute the data also benefit from the models. Consider for instance a recommender system (e.g. Spotify, Inst agram or YouTube), a health care app that predicts the risk for some disease, or a service built by pooling data across enterprises. In this work we propose a f ramework to study this value exchange, i.e., we model and measure contributions (outflows), benefits (inflows) and the balance between contributions and benefit s (the degree of reciprocity). We show theoretically, and via experiments that u nder certain distributional assumptions, some classes of models are approximatel y reciprocal. These results only scratch the surface; we conclude with several o pen directions.
****

When Personalization Harms Performance: Reconsidering the Use of Group Attribute s in Prediction
Vinith Menon Suriyakumar, Marzyeh Ghassemi, Berk Ustun
Machine learning models are often personalized with categorical attributes that define groups. In this work, we show that personalization with group attributes can inadvertently reduce performance at a group level – i.e., groups may receive unnecessarily inaccurate predictions by sharing their personal characteristics. We present formal conditions to ensure the fair use of group attributes in a pr ediction task, and describe how they can be checked by training one additional m odel. We characterize how fair use conditions be violated due to standard practi ces in model development, and study the prevalence of fair use violations in cli nical prediction tasks. Our results show that personalization often fails to pro duce a tailored performance gain for every group who reports personal data, and underscore the need to evaluate fair use when personalizing models with characte ristics that are protected, sensitive, self-reported, or costly to acquire.
****

Tuning Computer Vision Models With Task Rewards
André Susano Pinto, Alexander Kolesnikov, Yuge Shi, Lucas Beyer, Xiaohua Zhai
Misalignment between model predictions and intended usage can be detrimental for the deployment of computer vision models. The issue is exacerbated when the tas k involves complex structured outputs, as it becomes harder to design procedures which address this misalignment. In natural language processing, this is often addressed using reinforcement learning techniques that align models with a task reward. We adopt this approach and show its surprising effectiveness to improve generic models pretrained to imitate example outputs across multiple computer vi sion tasks, such as object detection, panoptic segmentation, colorization and im age captioning. We believe this approach has the potential to be widely useful f or better aligning models with a diverse range of computer vision tasks.
****

Beyond Exponentially Fast Mixing in Average-Reward Reinforcement Learning via Mu lti-Level Monte Carlo Actor-Critic
Wesley A Suttle, Amrit Bedi, Bhrij Patel, Brian M. Sadler, Alec Koppel, Dinesh M anocha
Many existing reinforcement learning (RL) methods employ stochastic gradient ite ration on the back end, whose stability hinges upon a hypothesis that the data-g enerating process mixes exponentially fast with a rate parameter that appears in the step-size selection. Unfortunately, this assumption is violated for large s tate spaces or settings with sparse rewards, and the mixing time is unknown, mak ing the step size inoperable. In this work, we propose an RL methodology attuned to the mixing time by employing a multi-level Monte Carlo estimator for the cri

tic, the actor, and the average reward embedded within an actor-critic (AC) algo rithm. This method, which we call Multi-level Actor-Critic (MAC), is developed s pecifically for infinite-horizon average-reward settings and neither relies on o racle knowledge of the mixing time in its parameter selection nor assumes its ex ponential decay; it is therefore readily applicable to applications with slower mixing times. Nonetheless, it achieves a convergence rate comparable to SOTA act or-critic algorithms. We experimentally show that these alleviated restrictions on the technical conditions required for stability translate to superior perform ance in practice for RL problems with sparse rewards.
****

Tight and fast generalization error bound of graph embedding in metric space
Atsushi Suzuki, Atsushi Nitanda, Taiji Suzuki, Jing Wang, Feng Tian, Kenji Yaman ishi
Recent studies have experimentally shown that we can achieve in non-Euclidean me tric space effective and efficient graph embedding, which aims to obtain the ver tices' representations reflecting the graph's structure in the metric space. Spe cifically, graph embedding in hyperbolic space has experimentally succeeded in e mbedding graphs with hierarchical-tree structure, e.g., data in natural language s, social networks, and knowledge bases. However, recent theoretical analyses ha ve shown a much higher upper bound on non-Euclidean graph embedding's generaliza tion error than Euclidean one's, where a high generalization error indicates tha t the incompleteness and noise in the data can significantly damage learning per formance. It implies that the existing bound cannot guarantee the success of gra ph embedding in non-Euclidean metric space in a practical training data size, wh ich can prevent non-Euclidean graph embedding's application in real problems. Th is paper provides a novel upper bound of graph embedding's generalization error by evaluating the local Rademacher complexity of the model as a function set of the distances of representation couples. Our bound clarifies that the performanc e of graph embedding in non-Euclidean metric space, including hyperbolic space, is better than the existing upper bounds suggest. Specifically, our new upper bo und is polynomial in the metric space's geometric radius R and can be $O(\frac {1}{S})$ at the fastest, where S is the training data size. Our bound is signi ficantly tighter and faster than the existing one, which can be exponential to $ R$ and $O(\frac{1}{\sqrt{S}})$ at the fastest. Specific calculations on example cases show that graph embedding in non-Euclidean metric space can outperform tha t in Euclidean space with much smaller training data than the existing bound has suggested.
****

Proximal Causal Learning of Conditional Average Treatment Effects
Erik Sverdrup, Yifan Cui
Efficiently and flexibly estimating treatment effect heterogeneity is an importa nt task in a wide variety of settings ranging from medicine to marketing, and th ere are a considerable number of promising conditional average treatment effect estimators currently available. These, however, typically rely on the assumption that the measured covariates are enough to justify conditional exchangeability. We propose the P-learner, motivated by the R- and DR-learner, a tailored two-st age loss function for learning heterogeneous treatment effects in settings where exchangeability given observed covariates is an implausible assumption, and we wish to rely on proxy variables for causal inference. Our proposed estimator can be implemented by off-the-shelf loss-minimizing machine learning methods, which in the case of kernel regression satisfies an oracle bound on the estimated err or as long as the nuisance components are estimated reasonably well.
****

Inverse Reinforcement Learning without Reinforcement Learning
Gokul Swamy, David Wu, Sanjiban Choudhury, Drew Bagnell, Steven Wu
Inverse Reinforcement Learning (IRL) is a powerful set of techniques for imitati on learning that aims to learn a reward function that rationalizes expert demons trations. Unfortunately, traditional IRL methods suffer from a computational wea kness: they require repeatedly solving a hard reinforcement learning (RL) proble m as a subroutine. This is counter-intuitive from the viewpoint of reductions: w

e have reduced the easier problem of imitation learning to repeatedly solving th e harder problem of RL. Another thread of work has proved that access to the sid e-information of the distribution of states where a strong policy spends time ca n dramatically reduce the sample and computational complexities of solving an RL problem. In this work, we demonstrate for the first time a more informed imitat ion learning reduction where we utilize the state distribution of the expert to alleviate the global exploration component of the RL subroutine, providing an ex ponential speedup in theory. In practice, we find that we are able to significan tly speed up the prior art on continuous control tasks.
****

Von Mises Mixture Distributions for Molecular Conformation Generation
Kirk Swanson, Jake Lawrence Williams, Eric M Jonas
Molecules are frequently represented as graphs, but the underlying 3D molecular geometry (the locations of the atoms) ultimately determines most molecular prope rties. However, most molecules are not static and at room temperature adopt a wi de variety of geometries or $\textit{conformations}$. The resulting distribution on geometries $p(x)$ is known as the Boltzmann distribution, and many molecular properties are expectations computed under this distribution. Generating accura te samples from the Boltzmann distribution is therefore essential for computing these expectations accurately. Traditional sampling-based methods are computatio nally expensive, and most recent machine learning-based methods have focused on identifying $\textit{modes}$ in this distribution rather than generating true $\ textit{samples}$. Generating such samples requires capturing conformational vari ability, and it has been widely recognized that the majority of conformational v ariability in molecules arises from rotatable bonds. In this work, we present Vo nMisesNet, a new graph neural network that captures conformational variability v ia a variational approximation of rotatable bond torsion angles as a mixture of von Mises distributions. We demonstrate that VonMisesNet can generate conformati ons for arbitrary molecules in a way that is both physically accurate with respe ct to the Boltzmann distribution and orders of magnitude faster than existing sa mpling methods.
****

Optimal randomized multilevel Monte Carlo for repeatedly nested expectations
Yasa Syed, Guanyang Wang
The estimation of repeatedly nested expectations is a challenging task that aris es in many real-world systems. However, existing methods generally suffer from h igh computational costs when the number of nestings becomes large. Fix any non-n egative integer D for the total number of nestings. Standard Monte Carlo metho ds typically cost at least $\mathcal{O}(\varepsilon^{-(2+D)})$ and sometimes $\m athcal {O}(\varepsilon^{-2(1+D)})$ to obtain an estimator up to ε-er ror. More advanced methods, such as multilevel Monte Carlo, currently only exist for $D = 1$. In this paper, we propose a novel Monte Carlo estimator called $\m athsf{READ}$, which stands for “Recursive Estimator for Arbitrary Depth.” Our es timator has an optimal computational cost of $\mathcal{O}(\varepsilon^{-2})$ for every fixed D under suitable assumptions, and a nearly optimal computational cost of $\mathcal{O}(\varepsilon^{-2(1 + \delta)})$ for any $0 < \delta < \frac1 2$ under much more general assumptions. Our estimator is also unbiased, which ma kes it easy to parallelize. The key ingredients in our construction are an obser vation of the problem’s recursive structure and the recursive use of the randomi zed multilevel Monte Carlo method.
****

Adaptive Coordination in Social Embodied Rearrangement
Andrew Szot, Unnat Jain, Dhruv Batra, Zsolt Kira, Ruta Desai, Akshara Rai
We present the task of "Social Rearrangement", consisting of cooperative everyda y tasks like setting up the dinner table, tidying a house or unpacking groceries in a simulated multi-agent environment. In Social Rearrangement, two robots coo rdinate to complete a long-horizon task, using onboard sensing and egocentric ob servations, and no privileged information about the environment. We study zero-s hot coordination (ZSC) in this task, where an agent collaborates with a new part ner, emulating a scenario where a robot collaborates with a new human partner. P

rior ZSC approaches struggle to generalize in our complex and visually rich setting, and on further analysis, we find that they fail to generate diverse coordination behaviors at training time. To counter this, we propose Behavior Diversity Play (BDP), a novel ZSC approach that encourages diversity through a discriminability objective. Our results demonstrate that BDP learns adaptive agents that can tackle visual coordination, and zero-shot generalize to new partners in unseen environments, achieving 35% higher success and 32% higher efficiency compared to baselines.

****

MG-GNN: Multigrid Graph Neural Networks for Learning Multilevel Domain Decomposition Methods

Ali Taghibakhshi, Nicolas Nytko, Tareq Uz Zaman, Scott Maclachlan, Luke Olson, Matthew West

Domain decomposition methods (DDMs) are popular solvers for discretized systems of partial differential equations (PDEs), with one-level and multilevel variants. These solvers rely on several algorithmic and mathematical parameters, prescribing overlap, subdomain boundary conditions, and other properties of the DDM. While some work has been done on optimizing these parameters, it has mostly focused on the one-level setting or special cases such as structured-grid discretizations with regular subdomain construction. In this paper, we propose multigrid graph neural networks (MG-GNN), a novel GNN architecture for learning optimized parameters in two-level DDMs. We train MG-GNN using a new unsupervised loss function, enabling effective training on small problems that yields robust performance on unstructured grids that are orders of magnitude larger than those in the training set. We show that MG-GNN outperforms popular hierarchical graph network architectures for this optimization and that our proposed loss function is critical to achieving this improved performance.

****

Learning Mixtures of Gaussians with Censored Data

Wai Ming Tai, Bryon Aragam

We study the problem of learning mixtures of Gaussians with censored data. Statistical learning with censored data is a classical problem, with numerous practical applications, however, finite-sample guarantees for even simple latent variable models such as Gaussian mixtures are missing. Formally, we are given censored data from a mixture of univariate Gaussians $ \sum_{i=1}^k w_i \mathcal{N}(\mu_i,\sigma^2), $ i.e. the sample is observed only if it lies inside a set S. The goal is to learn the weights w_i and the means μ_i. We propose an algorithm that takes only $\frac{1}{\varepsilon^{O(k)}}$ samples to estimate the weights w_i and the means μ_i within ε error.

****

Approximation and Estimation Ability of Transformers for Sequence-to-Sequence Functions with Infinite Dimensional Input

Shokichi Takakura, Taiji Suzuki

Despite the great success of Transformer networks in various applications such as natural language processing and computer vision, their theoretical aspects are not well understood. In this paper, we study the approximation and estimation ability of Transformers as sequence-to-sequence functions with infinite dimensional inputs. Although inputs and outputs are both infinite dimensional, we show that when the target function has anisotropic smoothness, Transformers can avoid the curse of dimensionality due to their feature extraction ability and parameter sharing property. In addition, we show that even if the smoothness changes depending on each input, Transformers can estimate the importance of features for each input and extract important features dynamically. Then, we proved that Transformers achieve similar convergence rate as in the case of the fixed smoothness. Our theoretical results support the practical success of Transformers for high dimensional data.

****

Learning Neural PDE Solvers with Parameter-Guided Channel Attention

Makoto Takamoto, Francesco Alesiani, Mathias Niepert

Scientific Machine Learning (SciML) is concerned with the development of learned

emulators of physical systems governed by partial differential equations (PDE). In application domains such as weather forecasting, molecular dynamics, and inverse design, ML-based surrogate models are increasingly used to augment or replace inefficient and often non-differentiable numerical simulation algorithms. While a number of ML-based methods for approximating the solutions of PDEs have been proposed in recent years, they typically do not adapt to the parameters of the PDEs, making it difficult to generalize to PDE parameters not seen during training. We propose a Channel Attention guided by PDE Parameter Embeddings (CAPE) component for neural surrogate models and a simple yet effective curriculum learning strategy. The CAPE module can be combined with any neural PDE solvers allowing them to adapt to unseen PDE parameters. The curriculum learning strategy provides a seamless transition between teacher-forcing and fully auto-regressive training. We compare CAPE in conjunction with the curriculum learning strategy using a PDE benchmark and obtain consistent and significant improvements over the baseline models. The experiments also show several advantages of CAPE, such as its increased ability to generalize to unseen PDE parameters without large increases inference time and parameter count. An implementation of the method and experiments are available at https://anonymous.4open.science/r/CAPE-ML4Sci-145B.

****

Contextual Conservative Interleaving Bandits

Kei Takemura

The performance of a bandit algorithm is usually measured by the cumulative rewards of the actions chosen by the algorithm. However, in many real-world applications, the rewards in each round should be good enough for reasons such as safety and fairness. In this paper, we investigate the contextual conservative interleaving bandit problem, which has a performance constraint that requires the chosen actions to be not much worse than given baseline actions in each round. This work is the first to simultaneously consider the following practical situations: (1) multiple actions are chosen in a round, (2) the feature vectors associated with given actions depend on the round, and (3) the performance constraints in each round that depend only on the actions chosen in that round. We propose a meta-algorithm, Greedy on Confidence Widths (GCW), that satisfies the performance constraints with high probability. GCW uses a standard bandit algorithm and achieves minimax optimal regret up to logarithmic factors if the algorithm used is also minimax optimal. We improve the existing analyses for the C^{2}UCB algorithm and the Thompson sampling to combine with GCW. We show that these algorithms achieve near-optimal regret when the feasible sets of given actions are the bases of a matroid. Our numerical experiments on a real-world dataset demonstrate that GCW with the standard bandit algorithms efficiently improves performance while satisfying the performance constraints.

****

Randomized Gaussian Process Upper Confidence Bound with Tighter Bayesian Regret Bounds

Shion Takeno, Yu Inatsu, Masayuki Karasuyama

Gaussian process upper confidence bound (GP-UCB) is a theoretically promising approach for black-box optimization; however, the confidence parameter β is considerably large in the theorem and chosen heuristically in practice. Then, randomized GP-UCB (RGP-UCB) uses a randomized confidence parameter, which follows the Gamma distribution, to mitigate the impact of manually specifying β. This study first generalizes the regret analysis of RGP-UCB to a wider class of distributions, including the Gamma distribution. Furthermore, we propose improved RGP-UCB (IRGP-UCB) based on a two-parameter exponential distribution, which achieves tighter Bayesian regret bounds. IRGP-UCB does not require an increase in the confidence parameter in terms of the number of iterations, which avoids over-exploration in the later iterations. Finally, we demonstrate the effectiveness of IRGP-UCB through extensive experiments.

****

Towards Practical Preferential Bayesian Optimization with Skew Gaussian Processes

Shion Takeno, Masahiro Nomura, Masayuki Karasuyama

We study preferential Bayesian optimization (BO) where reliable feedback is limi
ted to pairwise comparison called duels. An important challenge in preferential
BO, which uses the preferential Gaussian process (GP) model to represent flexibl
e preference structure, is that the posterior distribution is a computationally
intractable skew GP. The most widely used approach for preferential BO is Gaussi
an approximation, which ignores the skewness of the true posterior. Alternativel
y, Markov chain Monte Carlo (MCMC) based preferential BO is also proposed. In th
is work, we first verify the accuracy of Gaussian approximation, from which we r
eveal the critical problem that the predictive probability of duels can be inacc
urate. This observation motivates us to improve the MCMC-based estimation for sk
ew GP, for which we show the practical efficiency of Gibbs sampling and derive t
he low variance MC estimator. However, the computational time of MCMC can still
be a bottleneck in practice. Towards building a more practical preferential BO,
we develop a new method that achieves both high computational efficiency and low
 sample complexity, and then demonstrate its effectiveness through extensive num
erical experiments.
****

Robust Explanation for Free or At the Cost of Faithfulness

Zeren Tan, Yang Tian

Devoted to interpreting the explicit behaviors of machine learning models, expla
nation methods can identify implicit characteristics of models to improve trustw
orthiness. However, explanation methods are shown as vulnerable to adversarial p
erturbations, implying security concerns in high-stakes domains. In this paper,
we investigate when robust explanations are necessary and what they cost. We pro
ve that the robustness of explanations is determined by the robustness of the mo
del to be explained. Therefore, we can have robust explanations for free for a r
obust model. To have robust explanations for a non-robust model, composing the o
riginal model with a kernel is proved as an effective way that returns strictly
more robust explanations. Nevertheless, we argue that this also incurs a robustn
ess-faithfulness trade-off, i.e., contrary to common expectations, an explanatio
n method may also become less faithful when it becomes more robust. This argumen
t holds for any model. We are the first to introduce this trade-off and theoreti
cally prove its existence for SmoothGrad. Theoretical findings are verified by e
mpirical evidence on six state-of-the-art explanation methods and four backbones
.
****

Provably Invariant Learning without Domain Information

Xiaoyu Tan, Lin Yong, Shengyu Zhu, Chao Qu, Xihe Qiu, Xu Yinghui, Peng Cui, Yuan
 Qi

Typical machine learning applications always assume the data follows independent
 and identically distributed (IID) assumptions. In contrast, this assumption is
frequently violated in real-world circumstances, leading to the Out-of-Distribut
ion (OOD) generalization problem and a major drop in model robustness. To mitiga
te this issue, the invariant learning technique is leveraged to distinguish betw
een spurious features and invariant features among all input features and to tra
in the model purely on the basis of the invariant features. Numerous invariant l
earning strategies imply that the training data should contain domain informatio
n. Such information includes the environment index or auxiliary information acqu
ired from prior knowledge. However, acquiring these information is typically imp
ossible in practice. In this study, we present TIVA for environment-independent
invariance learning, which requires no environment-specific information in train
ing data. We discover and prove that, given certain mild data conditions, it is
possible to train an environment partitioning policy based on attributes that ar
e independent of the targets and then conduct invariant risk minimization. We ex
amine our method in comparison to other baseline methods, which demonstrate supe
rior performance and excellent robustness under OOD, using multiple benchmarks.
****

Auto-Differentiation of Relational Computations for Very Large Scale Machine Lea
rning

Yuxin Tang, Zhimin Ding, Dimitrije Jankov, Binhang Yuan, Daniel Bourgeois, Chris

Jermaine
The relational data model was designed to facilitate large-scale data management and analytics. We consider the problem of how to differentiate computations expressed relationally. We show experimentally that a relational engine running an auto-differentiated relational algorithm can easily scale to very large datasets, and is competitive with state-of-the-art, special-purpose systems for large-scale distributed machine learning.
****
Regret-Minimizing Double Oracle for Extensive-Form Games
Xiaohang Tang, Le Cong Dinh, Stephen Marcus Mcaleer, Yaodong Yang
By incorporating regret minimization, double oracle methods have demonstrated rapid convergence to Nash Equilibrium (NE) in normal-form games and extensive-form games, through algorithms such as online double oracle (ODO) and extensive-form double oracle (XDO), respectively. In this study, we further examine the theoretical convergence rate and sample complexity of such regret minimization-based double oracle methods, utilizing a unified framework called Regret-Minimizing Double Oracle. Based on this framework, we extend ODO to extensive-form games and determine its sample complexity. Moreover, we demonstrate that the sample complexity of XDO can be exponential in the number of information sets $|S|$, owing to the exponentially decaying stopping threshold of restricted games. To solve this problem, we propose the Periodic Double Oracle (PDO) method, which has the lowest sample complexity among regret minimization-based double oracle methods, being only polynomial in $|S|$. Empirical evaluations on multiple poker and board games show that PDO achieves significantly faster convergence than previous double oracle algorithms and reaches a competitive level with state-of-the-art regret minimization methods.
****
From Perception to Programs: Regularize, Overparameterize, and Amortize
Hao Tang, Kevin Ellis
We develop techniques for synthesizing neurosymbolic programs. Such programs mix discrete symbolic processing with continuous neural computation. We relax this mixed discrete/continuous problem and jointly learn all modules with gradient descent, and also incorporate amortized inference, overparameterization, and a differentiable strategy for penalizing lengthy programs. Collectedly this toolbox improves the stability of gradient-guided program search, and suggests ways of learning both how to parse continuous input into discrete abstractions, and how to process those abstractions via symbolic code.
****
Understanding Self-Predictive Learning for Reinforcement Learning
Yunhao Tang, Zhaohan Daniel Guo, Pierre Harvey Richemond, Bernardo Avila Pires, Yash Chandak, Remi Munos, Mark Rowland, Mohammad Gheshlaghi Azar, Charline Le Lan, Clare Lyle, András György, Shantanu Thakoor, Will Dabney, Bilal Piot, Daniele Calandriello, Michal Valko
We study the learning dynamics of self-predictive learning for reinforcement learning, a family of algorithms that learn representations by minimizing the prediction error of their own future latent representations. Despite its recent empirical success, such algorithms have an apparent defect: trivial representations (such as constants) minimize the prediction error, yet it is obviously undesirable to converge to such solutions. Our central insight is that careful designs of the optimization dynamics are critical to learning meaningful representations. We identify that a faster paced optimization of the predictor and semi-gradient updates on the representation, are crucial to preventing the representation collapse. Then in an idealized setup, we show self-predictive learning dynamics carries out spectral decomposition on the state transition matrix, effectively capturing information of the transition dynamics. Building on the theoretical insights, we propose bidirectional self-predictive learning, a novel self-predictive algorithm that learns two representations simultaneously. We examine the robustness of our theoretical insights with a number of small-scale experiments and showcase the promise of the novel representation learning algorithm with large-scale experiments.

****
DoMo-AC: Doubly Multi-step Off-policy Actor-Critic Algorithm
Yunhao Tang, Tadashi Kozuno, Mark Rowland, Anna Harutyunyan, Remi Munos, Bernardo Avila Pires, Michal Valko
Multi-step learning applies lookahead over multiple time steps and has proved valuable in policy evaluation settings. However, in the optimal control case, the impact of multi-step learning has been relatively limited despite a number of prior efforts. Fundamentally, this might be because multi-step policy improvements require operations that cannot be approximated by stochastic samples, hence hindering the widespread adoption of such methods in practice. To address such limitations, we introduce doubly multi-step off-policy VI (DoMo-VI), a novel oracle algorithm that combines multi-step policy improvements and policy evaluations. DoMo-VI enjoys guaranteed convergence speed-up to the optimal policy and is applicable in general off-policy learning settings. We then propose doubly multi-step off-policy actor-critic (DoMo-AC), a practical instantiation of the DoMo-VI algorithm. DoMo-AC introduces a bias-variance trade-off that ensures improved policy gradient estimates. When combined with the IMPALA architecture, DoMo-AC has showed improvements over the baseline algorithm on Atari-57 game benchmarks.
****
Towards Understanding Generalization of Graph Neural Networks
Huayi Tang, Yong Liu
Graph neural networks (GNNs) are widely used in machine learning for graph-structured data. Even though GNNs have achieved remarkable success in real-world applications, understanding their working mechanism in theory is still on primary stage. In this paper, we move towards this goal from the perspective of generalization. Specifically, with consideration of stochastic optimization, we establish high probability bounds of generalization gap and gradients for transductive learning algorithms. After that, we provide high probability bounds of generalization gap for popular GNNs and analyze the factors affecting their generalization capability. These theoretical results reveal how the network architecture impacts the generalization gap. Experiments on benchmark datasets validate the theoretical findings. Our results provide new insights into understanding generalization of GNNs.
****
Towards a better understanding of representation dynamics under TD-learning
Yunhao Tang, Remi Munos
TD-learning is a foundation reinforcement learning (RL) algorithm for value prediction. Critical to the accuracy of value predictions is the quality of state representations. In this work, we consider the question: how does end-to-end TD-learning impact the representation over time? Complementary to prior work, we provide a set of analysis that sheds further light on the representation dynamics under TD-learning. We first show that when the environments are reversible, end-to-end TD-learning strictly decreases the value approximation error over time. Under further assumptions on the environments, we can connect the representation dynamics with spectral decomposition over the transition matrix. This latter finding establishes fitting multiple value functions from randomly generated rewards as a useful auxiliary task for representation learning, as we empirically validate on both tabular and Atari game suites.
****
VA-learning as a more efficient alternative to Q-learning
Yunhao Tang, Remi Munos, Mark Rowland, Michal Valko
In reinforcement learning, the advantage function is critical for policy improvement, but is often extracted from a learned Q-function. A natural question is: Why not learn the advantage function directly? In this work, we introduce VA-learning, which directly learns advantage function and value function using bootstrapping, without explicit reference to Q-functions. VA-learning learns off-policy and enjoys similar theoretical guarantees as Q-learning. Thanks to the direct learning of advantage function and value function, VA-learning improves the sample efficiency over Q-learning both in tabular implementations and deep RL agents on Atari-57 games. We also identify a close connection between VA-learning and th

e dueling architecture, which partially explains why a simple architectural chan ge to DQN agents tends to improve performance.
****

Defects of Convolutional Decoder Networks in Frequency Representation
Ling Tang, Wen Shen, Zhanpeng Zhou, Yuefeng Chen, Quanshi Zhang
In this paper, we prove the representation defects of a cascaded convolutional d ecoder network, considering the capacity of representing different frequency com ponents of an input sample. We conduct the discrete Fourier transform on each ch annel of the feature map in an intermediate layer of the decoder network. Then, we extend the 2D circular convolution theorem to represent the forward and backw ard propagations through convolutional layers in the frequency domain. Based on this, we prove three defects in representing feature spectrums. First, we prove that the convolution operation, the zero-padding operation, and a set of other s ettings all make a convolutional decoder network more likely to weaken high-freq uency components. Second, we prove that the upsampling operation generates a fea ture spectrum, in which strong signals repetitively appear at certain frequencie s. Third, we prove that if the frequency components in the input sample and freq uency components in the target output for regression have a small shift, then th e decoder usually cannot be effectively learned.
****

Difference-in-Differences Meets Tree-based Methods: Heterogeneous Treatment Effe cts Estimation with Unmeasured Confounding
Caizhi Tang, Huiyuan Wang, Xinyu Li, Qing Cui, Longfei Li, Jun Zhou
This study considers the estimation of conditional causal effects in the presenc e of unmeasured confounding for a balanced panel with treatment imposed at the l ast time point. To address this, we combine Difference-in-differences (DiD) and tree-based methods and propose a new identification assumption that allows for t he violation of the (conditional) parallel trends assumption adopted by most exi sting DiD methods. Under this new assumption, we prove partial identifiability o f the conditional average treatment effect on the treated group (CATT). Our prop osed method estimates CATT through a tree-based causal approach, guided by a nov el splitting rule that avoids model misspecification and unnecessary auxiliary p arameter estimation. The splitting rule measures both the error of fitting obser ved data and the violation of conditional parallel trends simultaneously. We als o develop an ensemble of multiple trees via gradient boosting to further enhance performance. Experimental results on both synthetic and real-world datasets val idate the effectiveness of our proposed method.
****

End-to-end Training of Deep Boltzmann Machines by Unbiased Contrastive Divergenc e with Local Mode Initialization
Shohei Taniguchi, Masahiro Suzuki, Yusuke Iwasawa, Yutaka Matsuo
We address the problem of biased gradient estimation in deep Boltzmann machines (DBMs). The existing method to obtain an unbiased estimator uses a maximal coupl ing based on a Gibbs sampler, but when the state is high-dimensional, it takes a long time to converge. In this study, we propose to use a coupling based on the Metropolis-Hastings (MH) and to initialize the state around a local mode of the target distribution. Because of the propensity of MH to reject proposals, the c oupling tends to converge in only one step with a high probability, leading to h igh efficiency. We find that our method allows DBMs to be trained in an end-to-e nd fashion without greedy pretraining. We also propose some practical techniques to further improve the performance of DBMs. We empirically demonstrate that our training algorithm enables DBMs to show comparable generative performance to ot her deep generative models, achieving the FID score of 10.33 for MNIST.
****

POUF: Prompt-Oriented Unsupervised Fine-tuning for Large Pre-trained Models
Korawat Tanwisuth, Shujian Zhang, Huangjie Zheng, Pengcheng He, Mingyuan Zhou
Through prompting, large-scale pre-trained models have become more expressive an d powerful, gaining significant attention in recent years. Though these big mode ls have zero-shot capabilities, in general, labeled data are still required to a dapt them to downstream tasks. To overcome this critical limitation, we propose

an unsupervised fine-tuning framework to directly fine-tune the model or prompt on the unlabeled target data. We demonstrate how to apply our method to both language-augmented vision and masked-language models, by aligning the discrete distributions extracted from the prompts and target data. To verify our approach’s applicability, we conduct extensive experiments on image classification, sentiment analysis, and natural language inference tasks. Across 13 image-related tasks and 15 language-related ones, the proposed approach achieves consistent improvements over the baselines. PyTorch code is available at https://github.com/korawat-tanwisuth/POUF.

****

Dual Focal Loss for Calibration

Linwei Tao, Minjing Dong, Chang Xu

The use of deep neural networks in real-world applications require well-calibrated networks with confidence scores that accurately reflect the actual probability. However, it has been found that these networks often provide over-confident predictions, which leads to poor calibration. Recent efforts have sought to address this issue by focal loss to reduce over-confidence, but this approach can also lead to under-confident predictions. While different variants of focal loss have been explored, it is difficult to find a balance between over-confidence and under-confidence. In our work, we propose a new loss function by focusing on dual logits. Our method not only considers the ground truth logit, but also take into account the highest logit ranked after the ground truth logit. By maximizing the gap between these two logits, our proposed dual focal loss can achieve a better balance between over-confidence and under-confidence. We provide theoretical evidence to support our approach and demonstrate its effectiveness through evaluations on multiple models and datasets, where it achieves state-of-the-art performance. Code is available at https://github.com/Linwei94/DualFocalLoss

****

Abstract-to-Executable Trajectory Translation for One-Shot Task Generalization

Stone Tao, Xiaochen Li, Tongzhou Mu, Zhiao Huang, Yuzhe Qin, Hao Su

Training long-horizon robotic policies in complex physical environments is essential for many applications, such as robotic manipulation. However, learning a policy that can generalize to unseen tasks is challenging. In this work, we propose to achieve one-shot task generalization by decoupling plan generation and plan execution. Specifically, our method solves complex long-horizon tasks in three steps: build a paired abstract environment by simplifying geometry and physics, generate abstract trajectories, and solve the original task by an abstract-to-executable trajectory translator. In the abstract environment, complex dynamics such as physical manipulation are removed, making abstract trajectories easier to generate. However, this introduces a large domain gap between abstract trajectories and the actual executed trajectories as abstract trajectories lack low-level details and are not aligned frame-to-frame with the executed trajectory. In a manner reminiscent of language translation, our approach leverages a seq-to-seq model to overcome the large domain gap between the abstract and executable trajectories, enabling the low-level policy to follow the abstract trajectory. Experimental results on various unseen long-horizon tasks with different robot embodiments demonstrate the practicability of our methods to achieve one-shot task generalization.

****

Data Feedback Loops: Model-driven Amplification of Dataset Biases

Rohan Taori, Tatsunori Hashimoto

Datasets scraped from the internet have been critical to large-scale machine learning. Yet, its success puts the utility of future internet-derived datasets at potential risk, as model outputs begin to replace human annotations as a source of supervision. In this work, we formalize a system where interactions with one model are recorded as history and scraped as training data in the future. We then analyze its stability over time by tracking changes to a test-time bias statistic (e.g. gender bias of model predictions). We find that the degree of bias amplification is closely linked to whether the model’s outputs behave like samples from the training distribution, a behavior which we characterize and define as u

niform faithfulness. Experiments in three conditional prediction scenarios – ima ge classification, visual role-labeling, and language generation – demonstrate t hat models that exhibit a sampling-like behavior are more faithful and thus more stable. Based on this insight, we propose an intervention to help mitigate and stabilize unstable feedback systems.
****
Deep Regression Unlearning
Ayush Kumar Tarun, Vikram Singh Chundawat, Murari Mandal, Mohan Kankanhalli
With the introduction of data protection and privacy regulations, it has become crucial to remove the lineage of data on demand from a machine learning (ML) mod el. In the last few years, there have been notable developments in machine unlea rning to remove the information of certain training data efficiently and effecti vely from ML models. In this work, we explore unlearning for the regression prob lem, particularly in deep learning models. Unlearning in classification and simp le linear regression has been considerably investigated. However, unlearning in deep regression models largely remains an untouched problem till now. In this wo rk, we introduce deep regression unlearning methods that generalize well and are robust to privacy attacks. We propose the Blindspot unlearning method which use s a novel weight optimization process. A randomly initialized model, partially e xposed to the retain samples and a copy of the original model are used together to selectively imprint knowledge about the data that we wish to keep and scrub o ff the information of the data we wish to forget. We also propose a Gaussian fin e tuning method for regression unlearning. The existing unlearning metrics for c lassification are not directly applicable to regression unlearning. Therefore, w e adapt these metrics for the regression setting. We conduct regression unlearni ng experiments for computer vision, natural language processing and forecasting applications. Our methods show excellent performance for all these datasets acro ss all the metrics. Source code: https://github.com/ayu987/deep-regression-unlea rning
****
How to Trust Your Diffusion Model: A Convex Optimization Approach to Conformal R isk Control
Jacopo Teneggi, Matthew Tivnan, Web Stayman, Jeremias Sulam
Score-based generative modeling, informally referred to as diffusion models, con tinue to grow in popularity across several important domains and tasks. While th ey provide high-quality and diverse samples from empirical distributions, import ant questions remain on the reliability and trustworthiness of these sampling pr ocedures for their responsible use in critical scenarios. Conformal prediction i s a modern tool to construct finite-sample, distribution-free uncertainty guaran tees for any black-box predictor. In this work, we focus on image-to-image regre ssion tasks and we present a generalization of the Risk-Controlling Prediction S ets (RCPS) procedure, that we term K-RCPS, which allows to (i) provide entry wise calibrated intervals for future samples of any diffusion model, and (ii) control a certain notion of risk with respect to a ground truth image with minim al mean interval length. Differently from existing conformal risk control proced ures, ours relies on a novel convex optimization approach that allows for multid imensional risk control while provably minimizing the mean interval length. We i llustrate our approach on two real-world image denoising problems: on natural im ages of faces as well as on computed tomography (CT) scans of the abdomen, demon strating state of the art performance.
****
Concurrent Shuffle Differential Privacy Under Continual Observation
Jay Tenenbaum, Haim Kaplan, Yishay Mansour, Uri Stemmer
We introduce the concurrent shuffle model of differential privacy. In this model we have multiple concurrent shufflers permuting messages from different, possib ly overlapping, batches of users. Similarly to the standard (single) shuffler mo del, the privacy requirement is that the concatenation of all shuffled messages should be differentially private. We study the private continual summation probl em (a.k.a. the counter problem) and show that the concurrent shuffle model allow s for significantly improved error compared to a standard (single) shuffler mode

l. Specifically, we give a summation algorithm with error $\tilde{O}(n^{1/(2k+1)})$ with k concurrent shufflers on a sequence of length n. Furthermore, we prove that this bound is tight for any k, even if the algorithm can choose the sizes of the batches adaptively. For $k=\log n$ shufflers, the resulting error is polylogarithmic, much better than $\tilde{\Theta}(n^{1/3})$ which we show is the smallest possible with a single shuffler. We use our online summation algorithm to get algorithms with improved regret bounds for the contextual linear bandit problem. In particular we get optimal $\tilde{O}(\sqrt{n})$ regret with $k= \tilde{\Omega}(\log n)$ concurrent shufflers.

****

Finding Generalization Measures by Contrasting Signal and Noise

Jiaye Teng, Bohang Zhang, Ruichen Li, Haowei He, Yequan Wang, Yan Tian, Yang Yuan

Generalization is one of the most fundamental challenges in deep learning, aiming to predict model performances on unseen data. Empirically, such predictions usually rely on a validation set, while recent works showed that an unlabeled validation set also works. Without validation sets, it is extremely difficult to obtain non-vacuous generalization bounds, which leads to a weaker task of finding generalization measures that monotonically relate to generalization error. In this paper, we propose a new generalization measure REF Complexity (RElative Fitting degree between signal and noise), motivated by the intuition that a given model-algorithm pair may generalize well if it fits signal (e.g., true labels) fast while fitting noise (e.g., random labels) slowly. Empirically, REF Complexity monotonically relates to test accuracy in real-world datasets without accessing additional validation sets, achieving -0.988 correlation on CIFAR-10 and -0.960 correlation on CIFAR-100. We further theoretically verify the utility of REF Complexity under three different cases, including convex and smooth regimes with stochastic gradient descent, smooth regimes (not necessarily convex) with stochastic gradient Langevin dynamics, and linear regimes with gradient descent. The code is available at https://github.com/962086838/REF-complexity.

****

Reinforcement Learning with History Dependent Dynamic Contexts

Guy Tennenholtz, Nadav Merlis, Lior Shani, Martin Mladenov, Craig Boutilier

We introduce Dynamic Contextual Markov Decision Processes (DCMDPs), a novel reinforcement learning framework for history-dependent environments that generalizes the contextual MDP framework to handle non-Markov environments, where contexts change over time. We consider special cases of the model, with a focus on logistic DCMDPs, which break the exponential dependence on history length by leveraging aggregation functions to determine context transitions. This special structure allows us to derive an upper-confidence-bound style algorithm for which we establish regret bounds. Motivated by our theoretical results, we introduce a practical model-based algorithm for logistic DCMDPs that plans in a latent space and uses optimism over history-dependent features. We demonstrate the efficacy of our approach on a recommendation task (using MovieLens data) where user behavior dynamics evolve in response to recommendations.

****

PWSHAP: A Path-Wise Explanation Model for Targeted Variables

Lucile Ter-Minassian, Oscar Clivio, Karla Diazordaz, Robin J. Evans, Christopher C. Holmes

Predictive black-box models can exhibit high-accuracy but their opaque nature hinders their uptake in safety-critical deployment environments. Explanation methods (XAI) can provide confidence for decision-making through increased transparency. However, existing XAI methods are not tailored towards models in sensitive domains where one predictor is of special interest, such as a treatment effect in a clinical model, or ethnicity in policy models. We introduce Path-Wise Shapley effects (PWSHAP), a framework for assessing the targeted effect of a binary (e.g. treatment) variable from a complex outcome model. Our approach augments the predictive model with a user-defined directed acyclic graph (DAG). The method then uses the graph alongside on-manifold Shapley values to identify effects along causal pathways whilst maintaining robustness to adversarial attacks. We establi

sh error bounds for the identified path-wise Shapley effects and for Shapley values. We show PWSHAP can perform local bias and mediation analyses with faithfulness to the model. Further, if the targeted variable is randomised we can quantify local effect modification. We demonstrate the resolution, interpretability and true locality of our approach on examples and a real-world experiment.

****

On the Estimation of Gaussian Mixture Copula Models

Ashutosh Tewari

This paper revisits Gaussian Mixture Copula Model (GMCM), a more expressive alternative to the widely used Gaussian Mixture Model (GMM), with the goal to make its parameter estimation tractable. Both the Expectation Maximization and the direct Likelihood Maximization frameworks for GMCM have to grapple with a likelihood function that lacks a closed form. This has led to a few approximation schemes that alleviate the problem, nonetheless leaving the issue still unresolved. Additionally, past works have alluded to an additional challenge of parameter non-identifiability, but none has offered a rigorous treatment and a commensurate solution framework to overcome the same. This work offers solutions to each of these issues in an attempt to help GMCM realize its full potential. The source of non-identifiability is not only proven but also suitable priors are proposed that eliminate the problem. Additionally, an efficient numerical framework is proposed to evaluate the intractable likelihood function, while also providing its analytical derivatives. Finally, a view of GMCM as a series of bijective mappings from a base distribution is presented, which paves the way to synthesize GMCM using modern, probabilistic programming languages (PPLs). The main claims of this work are supported by empirical evidence gathered on synthetic and real-world data sets.

****

Target-Aware Generative Augmentations for Single-Shot Adaptation

Kowshik Thopalli, Rakshith Subramanyam, Pavan K. Turaga, Jayaraman J. Thiagarajan

In this paper, we address the problem of adapting models from a source domain to a target domain, a task that has become increasingly important due to the brittle generalization of deep neural networks. While several test-time adaptation techniques have emerged, they typically rely on synthetic toolbox data augmentations in cases of limited target data availability. We consider the challenging setting of single-shot adaptation and explore the design of augmentation strategies. We argue that augmentations utilized by existing methods are insufficient to handle large distribution shifts, and hence propose a new approach SiSTA, which first fine-tunes a generative model from the source domain using a single-shot target, and then employs novel sampling strategies for curating synthetic target data. Using experiments on a variety of benchmarks, distribution shifts and image corruptions, we find that SiSTA produces significantly improved generalization over existing baselines in face attribute detection and multi-class object recognition. Furthermore, SiSTA performs competitively to models obtained by training on larger target datasets. Our codes can be accessed at https://github.com/Rakshith-2905/SiSTA

****

ELSA: Efficient Label Shift Adaptation through the Lens of Semiparametric Models

Qinglong Tian, Xin Zhang, Jiwei Zhao

We study the domain adaptation problem with label shift in this work. Under the label shift context, the marginal distribution of the label varies across the training and testing datasets, while the conditional distribution of features given the label is the same. Traditional label shift adaptation methods either suffer from large estimation errors or require cumbersome post-prediction calibrations. To address these issues, we first propose a moment-matching framework for adapting the label shift based on the geometry of the influence function. Under such a framework, we propose a novel method named $\underline{\mathrm{E}}$fficient $\underline{\mathrm{L}}$abel $\underline{\mathrm{S}}$hift $\underline{\mathrm{A}}$daptation (ELSA), in which the adaptation weights can be estimated by solving linear systems. Theoretically, the ELSA estimator is $\sqrt{n}$-consistent (n

is the sample size of the source data) and asymptotically normal. Empirically, w
e show that ELSA can achieve state-of-the-art estimation performances without po
st-prediction calibrations, thus, gaining computational efficiency.
****

Spherical Inducing Features for Orthogonally-Decoupled Gaussian Processes
Louis C. Tiao, Vincent Dutordoir, Victor Picheny
Despite their many desirable properties, Gaussian processes (GPs) are often comp
ared unfavorably to deep neural networks (NNs) for lacking the ability to learn
representations. Recent efforts to bridge the gap between GPs and deep NNs have
yielded a new class of inter-domain variational GPs in which the inducing variab
les correspond to hidden units of a feedforward NN. In this work, we examine som
e practical issues associated with this approach and propose an extension that l
everages the orthogonal decomposition of GPs to mitigate these limitations. In p
articular, we introduce spherical inter-domain features to construct more flexib
le data-dependent basis functions for both the principal and orthogonal componen
ts of the GP approximation and show that incorporating NN activation features un
der this framework not only alleviates these shortcomings but is more scalable t
han alternative strategies. Experiments on multiple benchmark datasets demonstra
te the effectiveness of our approach.
****

Fast Rates for Maximum Entropy Exploration
Daniil Tiapkin, Denis Belomestny, Daniele Calandriello, Eric Moulines, Remi Muno
s, Alexey Naumov, Pierre Perrault, Yunhao Tang, Michal Valko, Pierre Menard
We address the challenge of exploration in reinforcement learning (RL) when the
agent operates in an unknown environment with sparse or no rewards. In this work
, we study the maximum entropy exploration problem of two different types. The f
irst type is visitation entropy maximization previously considered by Hazan et a
l. (2019) in the discounted setting. For this type of exploration, we propose a
game-theoretic algorithm that has $\widetilde{\mathcal{O}}(H^3S^2A/\varepsilon^2
)$ sample complexity thus improving the ε-dependence upon existing r
esults, where S is a number of states, A is a number of actions, H is an e
pisode length, and ε is a desired accuracy. The second type of entro
py we study is the trajectory entropy. This objective function is closely relate
d to the entropy-regularized MDPs, and we propose a simple algorithm that has a
sample complexity of order $\widetilde{\mathcal{O}}(\mathrm{poly}(S,A,H)/\vareps
ilon)$. Interestingly, it is the first theoretical result in RL literature that
establishes the potential statistical advantage of regularized MDPs for explorat
ion. Finally, we apply developed regularization techniques to reduce sample comp
lexity of visitation entropy maximization to $\widetilde{\mathcal{O}}(H^2SA/\var
epsilon^2)$, yielding a statistical separation between maximum entropy explorati
on and reward-free exploration.
****

Margin-based sampling in high dimensions: When being active is less efficient th
an staying passive
Alexandru Tifrea, Jacob Clarysse, Fanny Yang
It is widely believed that given the same labeling budget, active learning (AL)
algorithms like margin-based active learning achieve better predictive performan
ce than passive learning (PL), albeit at a higher computational cost. Recent emp
irical evidence suggests that this added cost might be in vain, as margin-based
AL can sometimes perform even worse than PL. While existing works offer differen
t explanations in the low-dimensional regime, this paper shows that the underlyi
ng mechanism is entirely different in high dimensions: we prove for logistic reg
ression that PL outperforms margin-based AL even for noiseless data and when usi
ng the Bayes optimal decision boundary for sampling. Insights from our proof ind
icate that this high-dimensional phenomenon is exacerbated when the separation b
etween the classes is small. We corroborate this intuition with experiments on 2
0 high-dimensional datasets spanning a diverse range of applications, from finan
ce and histology to chemistry and computer vision.
****

Differentiable Multi-Target Causal Bayesian Experimental Design

Panagiotis Tigas, Yashas Annadani, Desi R. Ivanova, Andrew Jesson, Yarin Gal, Adam Foster, Stefan Bauer
We introduce a gradient-based approach for the problem of Bayesian optimal experimental design to learn causal models in a batch setting — a critical component for causal discovery from finite data where interventions can be costly or risky. Existing methods rely on greedy approximations to construct a batch of experiments while using black-box methods to optimize over a single target-state pair to intervene with. In this work, we completely dispose of the black-box optimization techniques and greedy heuristics and instead propose a conceptually simple end-to-end gradient-based optimization procedure to acquire a set of optimal intervention target-value pairs. Such a procedure enables parameterization of the design space to efficiently optimize over a batch of multi-target-state interventions, a setting which has hitherto not been explored due to its complexity. We demonstrate that our proposed method outperforms baselines and existing acquisition strategies in both single-target and multi-target settings across a number of synthetic datasets.
****
PCA-based Multi-Task Learning: a Random Matrix Approach
Malik Tiomoko, Romain Couillet, Frederic Pascal
The article proposes and theoretically analyses a computationally efficient multi-task learning (MTL) extension of popular principal component analysis (PCA)-based supervised learning schemes. The analysis reveals that (i) by default, learning may dramatically fail by suffering from negative transfer, but that (ii) simple counter-measures on data labels avert negative transfer and necessarily result in improved performances. Supporting experiments on synthetic and real data benchmarks show that the proposed method achieves comparable performance with state-of-the-art MTL methods but at a significantly reduced computational cost.
****
Perturbation Analysis of Neural Collapse
Tom Tirer, Haoxiang Huang, Jonathan Niles-Weed
Training deep neural networks for classification often includes minimizing the training loss beyond the zero training error point. In this phase of training, a "neural collapse" behavior has been observed: the variability of features (outputs of the penultimate layer) of within-class samples decreases and the mean features of different classes approach a certain tight frame structure. Recent works analyze this behavior via idealized unconstrained features models where all the minimizers exhibit exact collapse. However, with practical networks and datasets, the features typically do not reach exact collapse, e.g., because deep layers cannot arbitrarily modify intermediate features that are far from being collapsed. In this paper, we propose a richer model that can capture this phenomenon by forcing the features to stay in the vicinity of a predefined features matrix (e.g., intermediate features). We explore the model in the small vicinity case via perturbation analysis and establish results that cannot be obtained by the previously studied models. For example, we prove reduction in the within-class variability of the optimized features compared to the predefined input features (via analyzing gradient flow on the "central-path" with minimal assumptions), analyze the minimizers in the near-collapse regime, and provide insights on the effect of regularization hyperparameters on the closeness to collapse. We support our theory with experiments in practical deep learning settings.
****
Overcoming Simplicity Bias in Deep Networks using a Feature Sieve
Rishabh Tiwari, Pradeep Shenoy
Simplicity bias is the concerning tendency of deep networks to over-depend on simple, weakly predictive features, to the exclusion of stronger, more complex features. This causes biased, incorrect model predictions in many real-world applications, exacerbated by incomplete training data containing spurious feature-label correlations. We propose a direct, interventional method for addressing simplicity bias in DNNs, which we call the feature sieve. We aim to automatically identify and suppress easily-computable spurious features in lower layers of the network, thereby allowing the higher network levels to extract and utilize richer,

more meaningful representations. We provide concrete evidence of this differenti al suppression & enhancement of relevant features on both controlled datasets an d real-world images, and report substantial gains on many real-world debiasing b enchmarks (11.4% relative gain on Imagenet-A; 3.2% on BAR, etc). Crucially, we o utperform many baselines that incorporate knowledge about known spurious or bias ed attributes, despite our method not using any such information. We believe tha t our feature sieve work opens up exciting new research directions in automated adversarial feature extraction & representation learning for deep networks.
****

Beyond In-Domain Scenarios: Robust Density-Aware Calibration
Christian Tomani, Futa Kai Waseda, Yuesong Shen, Daniel Cremers
Calibrating deep learning models to yield uncertainty-aware predictions is cruci al as deep neural networks get increasingly deployed in safety-critical applicat ions. While existing post-hoc calibration methods achieve impressive results on in-domain test datasets, they are limited by their inability to yield reliable u ncertainty estimates in domain-shift and out-of-domain (OOD) scenarios. We aim t o bridge this gap by proposing DAC, an accuracy-preserving as well as Density-Aw are Calibration method based on k-nearest-neighbors (KNN). In contrast to existi ng post-hoc methods, we utilize hidden layers of classifiers as a source for unc ertainty-related information and study their importance. We show that DAC is a g eneric method that can readily be combined with state-of-the-art post-hoc method s. DAC boosts the robustness of calibration performance in domain-shift and OOD, while maintaining excellent in-domain predictive uncertainty estimates. We demo nstrate that DAC leads to consistently better calibration across a large number of model architectures, datasets, and metrics. Additionally, we show that DAC im proves calibration substantially on recent large-scale neural networks pre-train ed on vast amounts of data.
****

Distribution Free Domain Generalization
Peifeng Tong, Wu Su, He Li, Jialin Ding, Zhan Haoxiang, Song Xi Chen
Accurate prediction of the out-of-distribution data is desired for a learning al gorithm. In domain generalization, training data from source domains tend to hav e different distributions from that of the target domain, while the target data are absence in the training process. We propose a Distribution Free Domain Gener alization (DFDG) procedure for classification by conducting standardization to a void the dominance of a few domains in the training process. The essence of the DFDG is its reformulating the cross domain/class discrepancy by pairwise two sam ple test statistics, and equally weights their importance or the covariance stru ctures to avoid dominant domain/class. A theoretical generalization bound is est ablished for the multi-class classification problem. The DFDG is shown to offer a superior performance in empirical studies with fewer hyperparameters, which me ans faster and easier implementation.
****

Extending Kernel PCA through Dualization: Sparsity, Robustness and Fast Algorith ms
Francesco Tonin, Alex Lambert, Panagiotis Patrinos, Johan Suykens
The goal of this paper is to revisit Kernel Principal Component Analysis (KPCA) through dualization of a difference of convex functions. This allows to naturall y extend KPCA to multiple objective functions and leads to efficient gradient-ba sed algorithms avoiding the expensive SVD of the Gram matrix. Particularly, we c onsider objective functions that can be written as Moreau envelopes, demonstrati ng how to promote robustness and sparsity within the same framework. The propose d method is evaluated on synthetic and realworld benchmarks, showing significant speedup in KPCA training time as well as highlighting the benefits in terms of robustness and sparsity.
****

Robust Weak Supervision with Variational Auto-Encoders
Francesco Tonolini, Nikolaos Aletras, Yunlong Jiao, Gabriella Kazai
Recent advances in weak supervision (WS) techniques allow to mitigate the enormo us cost and effort of human data annotation for supervised machine learning by a

utomating it using simple rule-based labelling functions (LFs). However, LFs nee d to be carefully designed, often requiring expert domain knowledge and extensiv e validation for existing WS methods to be effective. To tackle this, we propose the Weak Supervision Variational Auto-Encoder (WS-VAE), a novel framework that combines unsupervised representation learning and weak labelling to reduce the d ependence of WS on expert and manual engineering of LFs. Our technique learns fr om inputs and weak labels jointly to capture the input signals distribution with a latent space. The unsupervised representation component of the WS-VAE regular ises the inference of weak labels, while a specifically designed decoder allows the model to learn the relevance of LFs for each input. These unique features le ad to considerably improved robustness to the quality of LFs, compared to existi ng methods. An extensive empirical evaluation on a standard WS benchmark shows t hat our WS-VAE is competitive to state-of-the-art methods and substantially more robust to LF engineering.
****

Fully Bayesian Autoencoders with Latent Sparse Gaussian Processes
Ba-Hien Tran, Babak Shahbaba, Stephan Mandt, Maurizio Filippone
We present a fully Bayesian autoencoder model that treats both local latent vari ables and global decoder parameters in a Bayesian fashion. This approach allows for flexible priors and posterior approximations while keeping the inference cos ts low. To achieve this, we introduce an amortized MCMC approach by utilizing an implicit stochastic network to learn sampling from the posterior over local lat ent variables. Furthermore, we extend the model by incorporating a Sparse Gaussi an Process prior over the latent space, allowing for a fully Bayesian treatment of inducing points and kernel hyperparameters and leading to improved scalabilit y. Additionally, we enable Deep Gaussian Process priors on the latent space and the handling of missing data. We evaluate our model on a range of experiments fo cusing on dynamic representation learning and generative modeling, demonstrating the strong performance of our approach in comparison to existing methods that c ombine Gaussian Processes and autoencoders.
****

Discrete Key-Value Bottleneck
Frederik Träuble, Anirudh Goyal, Nasim Rahaman, Michael Curtis Mozer, Kenji Kawa guchi, Yoshua Bengio, Bernhard Schölkopf
Deep neural networks perform well on classification tasks where data streams are i.i.d. and labeled data is abundant. Challenges emerge with non-stationary trai ning data streams such as continual learning. One powerful approach that has add ressed this challenge involves pre-training of large encoders on volumes of read ily available data, followed by task-specific tuning. Given a new task, however, updating the weights of these encoders is challenging as a large number of weig hts needs to be fine-tuned, and as a result, they forget information about the p revious tasks. In the present work, we propose a model architecture to address t his issue, building upon a discrete bottleneck containing pairs of separate and learnable key-value codes. Our paradigm will be to encode; process the represent ation via a discrete bottleneck; and decode. Here, the input is fed to the pre-t rained encoder, the output of the encoder is used to select the nearest keys, an d the corresponding values are fed to the decoder to solve the current task. The model can only fetch and re-use a sparse number of these key-value pairs during inference, enabling localized and context-dependent model updates. We theoretic ally investigate the ability of the discrete key-value bottleneck to minimize th e effect of learning under distribution shifts and show that it reduces the comp lexity of the hypothesis class. We empirically verify the proposed method under challenging class-incremental learning scenarios and show that the proposed mode l — without any task boundaries — reduces catastrophic forgetting across a wide variety of pre-trained models, outperforming relevant baselines on this task.
****

Mimetic Initialization of Self-Attention Layers
Asher Trockman, J Zico Kolter
It is notoriously difficult to train Transformers on small datasets; typically, large pre-trained models are instead used as the starting point. We explore the

weights of such pre-trained Transformers (particularly for vision) to attempt to find reasons for this discrepancy. Surprisingly, we find that simply initializing the weights of self-attention layers so that they "look" more like their pre-trained counterparts allows us to train vanilla Transformers faster and to higher final accuracies, particularly on vision tasks such as CIFAR-10 and ImageNet classification, where we see gains in accuracy of over 5% and 4%, respectively. Our initialization scheme is closed form, learning-free, and very simple: we set the product of the query and key weights to be approximately the identity, and the product of the value and projection weights to approximately the negative identity. As this mimics the patterns we saw in pre-trained Transformers, we call the technique "mimetic initialization".

****

Representer Point Selection for Explaining Regularized High-dimensional Models

Che-Ping Tsai, Jiong Zhang, Hsiang-Fu Yu, Eli Chien, Cho-Jui Hsieh, Pradeep Kumar Ravikumar

We introduce a novel class of sample-based explanations we term high-dimensional representers, that can be used to explain the predictions of a regularized high-dimensional model in terms of importance weights for each of the training samples. Our workhorse is a novel representer theorem for general regularized high-dimensional models, which decomposes the model prediction in terms of contributions from each of the training samples: with positive (negative) values corresponding to positive (negative) impact training samples to the model’s prediction. We derive consequences for the canonical instances of ℓ_1 regularized sparse models and nuclear norm regularized low-rank models. As a case study, we further investigate the application of low-rank models in the context of collaborative filtering, where we instantiate high-dimensional representers for specific popular classes of models. Finally, we study the empirical performance of our proposed methods on three real-world binary classification datasets and two recommender system datasets. We also showcase the utility of high-dimensional representers in explaining model recommendations.

****

Expected Gradients of Maxout Networks and Consequences to Parameter Initialization

Hanna Tseran, Guido Montufar

We study the gradients of a maxout network with respect to inputs and parameters and obtain bounds for the moments depending on the architecture and the parameter distribution. We observe that the distribution of the input-output Jacobian depends on the input, which complicates a stable parameter initialization. Based on the moments of the gradients, we formulate parameter initialization strategies that avoid vanishing and exploding gradients in wide networks. Experiments with deep fully-connected and convolutional networks show that this strategy improves SGD and Adam training of deep maxout networks. In addition, we obtain refined bounds on the expected number of linear regions, results on the expected curve length distortion, and results on the NTK.

****

Provable Data Subset Selection For Efficient Neural Networks Training

Murad Tukan, Samson Zhou, Alaa Maalouf, Daniela Rus, Vladimir Braverman, Dan Feldman

Radial basis function neural networks (RBFNN) are well-known for their capability to approximate any continuous function on a closed bounded set with arbitrary precision given enough hidden neurons. In this paper, we introduce the first algorithm to construct coresets for RBFNNs, i.e., small weighted subsets that approximate the loss of the input data on any radial basis function network and thus approximate any function defined by an RBFNN on the larger input data. In particular, we construct coresets for radial basis and Laplacian loss functions. We then use our coresets to obtain a provable data subset selection algorithm for training deep neural networks. Since our coresets approximate every function, they also approximate the gradient of each weight in a neural network, which is a particular function on the input. We then perform empirical evaluations on function approximation and dataset subset selection on popular network architectures and

data sets, demonstrating the efficacy and accuracy of our coreset construction.
****
Jump-Start Reinforcement Learning
Ikechukwu Uchendu, Ted Xiao, Yao Lu, Banghua Zhu, Mengyuan Yan, Joséphine Simon, Matthew Bennice, Chuyuan Fu, Cong Ma, Jiantao Jiao, Sergey Levine, Karol Hausman
Reinforcement learning (RL) provides a theoretical framework for continuously improving an agent’s behavior via trial and error. However, efficiently learning policies from scratch can be very difficult, particularly for tasks that present exploration challenges. In such settings, it might be desirable to initialize RL with an existing policy, offline data, or demonstrations. However, naively performing such initialization in RL often works poorly, especially for value-based methods. In this paper, we present a meta algorithm that can use offline data, demonstrations, or a pre-existing policy to initialize an RL policy, and is compatible with any RL approach. In particular, we propose Jump-Start Reinforcement Learning (JSRL), an algorithm that employs two policies to solve tasks: a guide-policy, and an exploration-policy. By using the guide-policy to form a curriculum of starting states for the exploration-policy, we are able to efficiently improve performance on a set of simulated robotic tasks. We show via experiments that it is able to significantly outperform existing imitation and reinforcement learning algorithms, particularly in the small-data regime. In addition, we provide an upper bound on the sample complexity of JSRL and show that with the help of a guide-policy, one can improve the sample complexity for non-optimism exploration methods from exponential in horizon to polynomial.
****
Submodular Order Functions and Assortment Optimization
Rajan Udwani
We define a new class of set functions that in addition to being monotone and subadditive, also admit a very limited form of submodularity defined over a permutation of the ground set. We refer to this permutation as a submodular order. We give fast algorithms with strong approximation guarantees for maximizing submodular order functions under a variety of constraints. Applying this new notion to the problem of constrained assortment optimization in fundamental choice models, we obtain new algorithms that are both faster and have stronger approximation guarantees (in some cases, first algorithm with constant factor guarantee). We also show an intriguing connection to the maximization of monotone submodular functions in the streaming model, where we recover best known approximation guarantees as a corollary of our results.
****
Computationally Efficient PAC RL in POMDPs with Latent Determinism and Conditional Embeddings
Masatoshi Uehara, Ayush Sekhari, Jason D. Lee, Nathan Kallus, Wen Sun
We study reinforcement learning with function approximation for large-scale Partially Observable Markov Decision Processes (POMDPs) where the state space and observation space are large or even continuous. Particularly, we consider Hilbert space embeddings of POMDP where the feature of latent states and the feature of observations admit a conditional Hilbert space embedding of the observation emission process, and the latent state transition is deterministic. Under the function approximation setup where the optimal latent state-action Q-function is linear in the state feature, and the optimal Q-function has a gap in actions, we provide a computationally and statistically efficient algorithm for finding the exact optimal policy. We show our algorithm’s computational and statistical complexities scale polynomially with respect to the horizon and the intrinsic dimension of the feature on the observation space. Furthermore, we show both the deterministic latent transitions and gap assumptions are necessary to avoid statistical complexity exponential in horizon or dimension. Since our guarantee does not have an explicit dependence on the size of the state and observation spaces, our algorithm provably scales to large-scale POMDPs.
****
From Adaptive Query Release to Machine Unlearning

Enayat Ullah, Raman Arora
We formalize the problem of machine unlearning as design of efficient unlearning algorithms corresponding to learning algorithms which perform a selection of adaptive queries from structured query classes. We give efficient unlearning algorithms for linear and prefix-sum query classes. As applications, we show that unlearning in many problems, in particular, stochastic convex optimization (SCO), can be reduced to the above, yielding improved guarantees for the problem. In particular, for smooth Lipschitz losses and any $\rho>0$, our results yield an unlearning algorithm with excess population risk of $\tilde O\big(\frac{1}{\sqrt{n}} +\frac{\sqrt{d}}{n\rho}\big)$ with unlearning query (gradient) complexity $\tilde O(\rho \cdot \text{Retraining Complexity})$, where d is the model dimensionality and n is the initial number of samples. For non-smooth Lipschitz losses, we give an unlearning algorithm with excess population risk $\tilde O\big(\frac{1}{\sqrt{n}}+\big(\frac{\sqrt{d}}{n\rho}\big)^{1/2}\big)$ with the same unlearning query (gradient) complexity. Furthermore, in the special case of Generalized Linear Models (GLMs), such as those in linear and logistic regression, we get dimension-independent rates of $\tilde O\big(\frac{1}{\sqrt{n}} +\frac{1}{(n\rho)^{2/3}}\big)$ and $\tilde O\big(\frac{1}{\sqrt{n}} +\frac{1}{(n\rho)^{1/3}}\big)$ for smooth Lipschitz and non-smooth Lipschitz losses respectively. Finally, we give generalizations of the above from one unlearning request to dynamic streams consisting of insertions and deletions.
****

Private Federated Learning with Autotuned Compression
Enayat Ullah, Christopher A. Choquette-Choo, Peter Kairouz, Sewoong Oh
We propose new techniques for reducing communication in private federated learning without the need for setting or tuning compression rates. Our on-the-fly methods automatically adjust the compression rate based on the error induced during training, while maintaining provable privacy guarantees through the use of secure aggregation and differential privacy. Our techniques are provably instance-optimal for mean estimation, meaning that they can adapt to the “hardness of the problem” with minimal interactivity. We demonstrate the effectiveness of our approach on real-world datasets by achieving favorable compression rates without the need for tuning.
****

The Monge Gap: A Regularizer to Learn All Transport Maps
Théo Uscidda, Marco Cuturi
Optimal transport (OT) theory has been used in machine learning to study and characterize maps that can push-forward efficiently a probability measure onto another. Recent works have drawn inspiration from Brenier’s theorem, which states that when the ground cost is the squared-Euclidean distance, the “best” map to morph a continuous measure in $\mathcal{P}(\mathbb{R}^d)$ into another must be the gradient of a convex function. To exploit that result, Makkuva et. al (2020); Korotin et. al (2020) consider maps $T=\nabla f_\theta$, where f_θ is an input convex neural network (ICNN), as defined by Amos et. al (2017), and fit θ with SGD using samples. Despite their mathematical elegance, fitting OT maps with ICNNs raises many challenges, due notably to the many constraints imposed on θ; the need to approximate the conjugate of f_θ; or the limitation that they only work for the squared-Euclidean cost. More generally, we question the relevance of using Brenier’s result, which only applies to densities, to constrain the architecture of candidate maps fitted on samples. Motivated by these limitations, we propose a radically different approach to estimating OT maps: Given a cost c and a reference measure ρ, we introduce a regularizer, the Monge gap $\mathcal{M}^c_{\rho}(T)$ of a map T. That gap quantifies how far a map T deviates from the ideal properties we expect from a c-OT map. In practice, we drop all architecture requirements for T and simply minimize a distance (e.g., the Sinkhorn divergence) between $T\sharp\mu$ and ν, regularized by $\mathcal{M}^c_\rho(T)$. We study $\mathcal{M}^c_{\rho}$ and show how our simple pipeline significantly outperforms other baselines in practice.
****

Semi-Dual Unbalanced Quadratic Optimal Transport: fast statistical rates and con

vergent algorithm.
Adrien Vacher, François-Xavier Vialard
In this paper, we derive a semi-dual formulation for the problem of unbalanced q uadratic optimal transport and we study its stability properties, namely we give upper and lower bounds for the Bregman divergence of the new objective that hol d globally. We observe that the new objective gains even more convexity than in the balanced case. We use this formulation to prove the first results on statist ical estimation of UOT potentials and we leverage the extra convexity to recover super-parametric rates. Interestingly, unlike in the balanced case, we do not r equire the potentials to be smooth. Then, use variable metric descent to solve t he semi-dual problem for which we prove convergence at a $1/k$ rate for strongly convex potentials and exponential convergence in the balanced case when potenti als are also smooth. We emphasize that our convergence results has an interest o n its own as it generalizes previous convergence results to non-equivalent metri cs. Last, we instantiate a proof-of-concept tractable version of our theoretical algorithm that we benchmark on a 2D experiment in the balanced case and on a me dium dimension synthetic experiment in the unbalanced case.
****
Random Grid Neural Processes for Parametric Partial Differential Equations
Arnaud Vadeboncoeur, Ieva Kazlauskaite, Yanni Papandreou, Fehmi Cirak, Mark Giro lami, Omer Deniz Akyildiz
We introduce a new class of spatially stochastic physics and data informed deep latent models for parametric partial differential equations (PDEs) which operate through scalable variational neural processes. We achieve this by assigning pro bability measures to the spatial domain, which allows us to treat collocation gr ids probabilistically as random variables to be marginalised out. Adapting this spatial statistics view, we solve forward and inverse problems for parametric PD Es in a way that leads to the construction of Gaussian process models of solutio n fields. The implementation of these random grids poses a unique set of challen ges for inverse physics informed deep learning frameworks and we propose a new a rchitecture called Grid Invariant Convolutional Networks (GICNets) to overcome t hese challenges. We further show how to incorporate noisy data in a principled m anner into our physics informed model to improve predictions for problems where data may be available but whose measurement location does not coincide with any fixed mesh or grid. The proposed method is tested on a nonlinear Poisson problem , Burgers equation, and Navier-Stokes equations, and we provide extensive numeri cal comparisons. We demonstrate significant computational advantages over curren t physics informed neural learning methods for parametric PDEs while improving t he predictive capabilities and flexibility of these models.
****
Delayed Feedback in Kernel Bandits
Sattar Vakili, Danyal Ahmed, Alberto Bernacchia, Ciara Pike-Burke
Black box optimisation of an unknown function from expensive and noisy evaluatio ns is a ubiquitous problem in machine learning, academic research and industrial production. An abstraction of the problem can be formulated as a kernel based b andit problem (also known as Bayesian optimisation), where a learner aims at opt imising a kernelized function through sequential noisy observations. The existin g work predominantly assumes feedback is immediately available; an assumption wh ich fails in many real world situations, including recommendation systems, clini cal trials and hyperparameter tuning. We consider a kernel bandit problem under stochastically delayed feedback, and propose an algorithm with $\tilde{\mathcal{ O}}\left(\sqrt{\Gamma_k(T) T}+\mathbb{E}[\tau]\right)$ regret, where T is the number of time steps, $\Gamma_k(T)$ is the maximum information gain of the kerne l with T observations, and τ is the delay random variable. This represent s a significant improvement over the state of the art regret bound of $\tilde{\m athcal{O}}\left(\Gamma_k(T)\sqrt{ T}+\mathbb{E}[\tau]\Gamma_k(T)\right)$ reporte d in (Verma et al., 2022). In particular, for very non-smooth kernels, the infor mation gain grows almost linearly in time, trivializing the existing results. We also validate our theoretical results with simulations.
****

Synthetic Data, Real Errors: How (Not) to Publish and Use Synthetic Data
Boris Van Breugel, Zhaozhi Qian, Mihaela Van Der Schaar
Generating synthetic data through generative models is gaining interest in the ML community and beyond, promising a future where datasets can be tailored to individual needs. Unfortunately, synthetic data is usually not perfect, resulting in potential errors in downstream tasks. In this work we explore how the generative process affects the downstream ML task. We show that the naive synthetic data approach—using synthetic data as if it is real—leads to downstream models and analyses that do not generalize well to real data. As a first step towards better ML in the synthetic data regime, we introduce Deep Generative Ensemble (DGE)—a framework inspired by Deep Ensembles that aims to implicitly approximate the posterior distribution over the generative process model parameters. DGE improves downstream model training, evaluation, and uncertainty quantification, vastly outperforming the naive approach on average. The largest improvements are achieved for minority classes and low-density regions of the original data, for which the generative uncertainty is largest.
****
Trading-Off Payments and Accuracy in Online Classification with Paid Stochastic Experts
Dirk Van Der Hoeven, Ciara Pike-Burke, Hao Qiu, Nicolò Cesa-Bianchi
We investigate online classification with paid stochastic experts. Here, before making their prediction, each expert must be paid. The amount that we pay each expert directly influences the accuracy of their prediction through some unknown Lipschitz “productivity” function. In each round, the learner must decide how much to pay each expert and then make a prediction. They incur a cost equal to a weighted sum of the prediction error and upfront payments for all experts. We introduce an online learning algorithm whose total cost after T rounds exceeds that of a predictor which knows the productivity of all experts in advance by at most $\mathcal{O}\big(K^2(\ln T)\sqrt{T}\big)$ where K is the number of experts. In order to achieve this result, we combine Lipschitz bandits and online classification with surrogate losses. These tools allow us to improve upon the bound of order $T^{2/3}$ one would obtain in the standard Lipschitz bandit setting. Our algorithm is empirically evaluated on synthetic data.
****
Causal Isotonic Calibration for Heterogeneous Treatment Effects
Lars Van Der Laan, Ernesto Ulloa-Perez, Marco Carone, Alex Luedtke
We propose causal isotonic calibration, a novel nonparametric method for calibrating predictors of heterogeneous treatment effects. Furthermore, we introduce cross-calibration, a data-efficient variant of calibration that eliminates the need for hold-out calibration sets. Cross-calibration leverages cross-fitted predictors and generates a single calibrated predictor using all available data. Under weak conditions that do not assume monotonicity, we establish that both causal isotonic calibration and cross-calibration achieve fast doubly-robust calibration rates, as long as either the propensity score or outcome regression is estimated accurately in a suitable sense. The proposed causal isotonic calibrator can be wrapped around any black-box learning algorithm, providing robust and distribution-free calibration guarantees while preserving predictive performance.
****
Accounting For Informative Sampling When Learning to Forecast Treatment Outcomes Over Time
Toon Vanderschueren, Alicia Curth, Wouter Verbeke, Mihaela Van Der Schaar
Machine learning (ML) holds great potential for accurately forecasting treatment outcomes over time, which could ultimately enable the adoption of more individualized treatment strategies in many practical applications. However, a significant challenge that has been largely overlooked by the ML literature on this topic is the presence of informative sampling in observational data. When instances are observed irregularly over time, sampling times are typically not random, but rather informative–depending on the instance’s characteristics, past outcomes, and administered treatments. In this work, we formalize informative sampling as a covariate shift problem and show that it can prohibit accurate estimation of tr

eatment outcomes if not properly accounted for. To overcome this challenge, we p
resent a general framework for learning treatment outcomes in the presence of in
formative sampling using inverse intensity-weighting, and propose a novel method
, TESAR-CDE, that instantiates this framework using Neural CDEs. Using a simulat
ion environment based on a clinical use case, we demonstrate the effectiveness o
f our approach in learning under informative sampling.
****

Best Arm Identification in Multi-Agent Multi-Armed Bandits
Filippo Vannella, Alexandre Proutiere, Jaeseong Jeong
We investigate the problem of best arm identification in Multi-Agent Multi-Armed
 Bandits (MAMABs) where the rewards are defined through a factor graph. The obje
ctive is to find an optimal global action with a prescribed level of confidence
and minimal sample complexity. We derive a tight instance-specific lower bound o
f the sample complexity and characterize the corresponding optimal sampling stra
tegy. Unfortunately, this bound is obtained by solving a combinatorial optimizat
ion problem with a number of variables and constraints exponentially growing wit
h the number of agents. We leverage Mean Field (MF) techniques to obtain, in a c
omputationally efficient manner, an approximation of the lower bound. The approx
imation scales at most as ρK^d (where ρ, K, and d denote the numb
er of factors in the graph, the number of possible actions per agent, and the ma
ximal degree of the factor graph). We devise MF-TaS (Mean-Field-Track-and-Stop),
 an algorithm whose sample complexity provably matches our approximated lower bo
und. We illustrate the performance of MF-TaS numerically using both synthetic an
d real-world experiments (e.g., to solve the antenna tilt optimization problem i
n radio communication networks).
****

Conditional Tree Matching for Inference-Time Adaptation of Tree Prediction Model
s
Harshit Varma, Abhijeet Awasthi, Sunita Sarawagi
We present CTreeOT, a convergent, differentiable algorithm for matching two tree
s when each tree is conditioned on some input. Such conditional tree matching is
 useful for light-weight, few-shot adaptation of tree prediction models without
parameter fine-tuning. CTreeOT includes an alignment algorithm that extends the
popular Sinkhorn algorithm for matching tree nodes while supporting constraints
on tree edges. The algorithm involves alternating between matrix rescaling and m
essage passing updates, and can be efficiently expressed as GPU tensor operation
s. The second part of CTreeOT is fine-grained relevance-based reweighting of nod
es that makes the match scores useful for prediction tasks. We demonstrate the u
sefulness of CTreeOT for cross-schema adaptation of Text-to-SQL, a popular seman
tic parsing task. We show that compared to state-of-the-art methods, we achieve
significant increase in adaptation accuracy.
****

Optimal LP Rounding and Linear-Time Approximation Algorithms for Clustering Edge
-Colored Hypergraphs
Nate Veldt
We study the approximability of an existing framework for clustering edge-colore
d hypergraphs, which is closely related to chromatic correlation clustering and
is motivated by machine learning and data mining applications where the goal is
to cluster a set of objects based on multiway interactions of different categori
es or types. We present improved approximation guarantees based on linear progra
mming, and show they are tight by proving a matching integrality gap. Our result
s also include new approximation hardness results, a combinatorial 2-approximati
on whose runtime is linear in the hypergraph size, and several new connections t
o well-studied objectives such as vertex cover and hypergraph multiway cut.
****

Fast $(1+\varepsilon)$-Approximation Algorithms for Binary Matrix Factorization
Ameya Velingker, Maximilian Vötsch, David Woodruff, Samson Zhou
We introduce efficient $(1+\varepsilon)$-approximation algorithms for the binary
 matrix factorization (BMF) problem, where the inputs are a matrix $\mathbf{A}\i
n\{0,1\}^{n\times d}$, a rank parameter $k>0$, as well as an accuracy parameter

$\varepsilon>0$, and the goal is to approximate $\mathbf{A}$ as a product of low-rank factors $\mathbf{U}\in\{0,1\}^{n\times k}$ and $\mathbf{V}\in\{0,1\}^{k\times d}$. Equivalently, we want to find $\mathbf{U}$ and $\mathbf{V}$ that minimize the Frobenius loss $\|\mathbf{U}\mathbf{V} - \mathbf{A}\|_F^2$. Before this work, the state-of-the-art for this problem was the approximation algorithm of Kumar et. al. [ICML 2019], which achieves a C-approximation for some constant $C \ge 576$. We give the first $(1+\varepsilon)$-approximation algorithm using running time singly exponential in k, where k is typically a small integer. Our techniques generalize to other common variants of the BMF problem, admitting bicriteria $(1+\varepsilon)$-approximation algorithms for L_p loss functions and the setting where matrix operations are performed in $\mathbb{F}_2$. Our approach can be implemented in standard big data models, such as the streaming or distributed models.

****

The Virtues of Laziness in Model-based RL: A Unified Objective and Algorithms

Anirudh Vemula, Yuda Song, Aarti Singh, Drew Bagnell, Sanjiban Choudhury

We propose a novel approach to addressing two fundamental challenges in Model-based Reinforcement Learning (MBRL): the computational expense of repeatedly finding a good policy in the learned model, and the objective mismatch between model fitting and policy computation. Our "lazy" method leverages a novel unified objective, Performance Difference via Advantage in Model, to capture the performance difference between the learned policy and expert policy under the true dynamics. This objective demonstrates that optimizing the expected policy advantage in the learned model under an exploration distribution is sufficient for policy computation, resulting in a significant boost in computational efficiency compared to traditional planning methods. Additionally, the unified objective uses a value moment matching term for model fitting, which is aligned with the model’s usage during policy computation. We present two no-regret algorithms to optimize the proposed objective, and demonstrate their statistical and computational gains compared to existing MBRL methods through simulated benchmarks.

****

Learning the Right Layers a Data-Driven Layer-Aggregation Strategy for Semi-Supervised Learning on Multilayer Graphs

Sara Venturini, Andrea Cristofari, Francesco Rinaldi, Francesco Tudisco

Clustering (or community detection) on multilayer graphs poses several additional complications with respect to standard graphs as different layers may be characterized by different structures and types of information. One of the major challenges is to establish the extent to which each layer contributes to the cluster assignment in order to effectively take advantage of the multilayer structure and improve upon the classification obtained using the individual layers or their union. However, making an informed a-priori assessment about the clustering information content of the layers can be very complicated. In this work, we assume a semi-supervised learning setting, where the class of a small percentage of nodes is initially provided, and we propose a parameter-free Laplacian-regularized model that learns an optimal nonlinear combination of the different layers from the available input labels. The learning algorithm is based on a Frank-Wolfe optimization scheme with inexact gradient, combined with a modified Label Propagation iteration. We provide a detailed convergence analysis of the algorithm and extensive experiments on synthetic and real-world datasets, showing that the proposed method compares favourably with a variety of baselines and outperforms each individual layer when used in isolation.

****

Multi-Environment Pretraining Enables Transfer to Action Limited Datasets

David Venuto, Sherry Yang, Pieter Abbeel, Doina Precup, Igor Mordatch, Ofir Nachum

Using massive datasets to train large-scale models has emerged as a dominant approach for broad generalization in natural language and vision applications. In reinforcement learning, however, a key challenge is that available data of sequential decision making is often not annotated with actions - for example, videos of game-play are much more available than sequences of frames paired with their l

ogged game controls. We propose to circumvent this challenge by combining large but sparsely-annotated datasets from a target environment of interest with fully-annotated datasets from various other source environments. Our method, Action Limited PreTraining (ALPT), leverages the generalization capabilities of inverse dynamics modelling (IDM) to label missing action data in the target environment. We show that utilizing even one additional environment dataset of labelled data during IDM pretraining gives rise to substantial improvements in generating action labels for unannotated sequences. We evaluate our method on benchmark game-playing environments and show that we can significantly improve game performance and generalization capability compared to other approaches, using annotated datasets equivalent to only 12 minutes of gameplay. Highlighting the power of IDM, we show that these benefits remain even when target and source environments share no common actions.

****

AbODE: Ab initio antibody design using conjoined ODEs

Yogesh Verma, Markus Heinonen, Vikas Garg

Antibodies are Y-shaped proteins that neutralize pathogens and constitute the core of our adaptive immune system. De novo generation of new antibodies that target specific antigens holds the key to accelerating vaccine discovery. However, this co-design of the amino acid sequence and the 3D structure subsumes and accentuates, some central challenges from multiple tasks including protein folding (sequence to structure), inverse folding (structure to sequence), and docking (binding). We strive to surmount these challenges with a new generative model AbODE that extends graph PDEs to accommodate both contextual information and external interactions. Unlike existing approaches, AbODE uses a single round of full-shot decoding, and elicits continuous differential attention that encapsulates, and evolves with, latent interactions within the antibody as well as those involving the antigen. We unravel fundamental connections between AbODE and temporal networks as well as graph-matching networks. The proposed model significantly outperforms existing methods on standard metrics across benchmarks.

****

TabLeak: Tabular Data Leakage in Federated Learning

Mark Vero, Mislav Balunovic, Dimitar Iliev Dimitrov, Martin Vechev

While federated learning (FL) promises to preserve privacy, recent works in the image and text domains have shown that training updates leak private client data. However, most high-stakes applications of FL (e.g., in healthcare and finance) use tabular data, where the risk of data leakage has not yet been explored. A successful attack for tabular data must address two key challenges unique to the domain: (i) obtaining a solution to a high-variance mixed discrete-continuous optimization problem, and (ii) enabling human assessment of the reconstruction as unlike for image and text data, direct human inspection is not possible. In this work we address these challenges and propose TabLeak, the first comprehensive reconstruction attack on tabular data. TabLeak is based on two key contributions: (i) a method which leverages a softmax relaxation and pooled ensembling to solve the optimization problem, and (ii) an entropy-based uncertainty quantification scheme to enable human assessment. We evaluate TabLeak on four tabular datasets for both FedSGD and FedAvg training protocols, and show that it successfully breaks several settings previously deemed safe. For instance, we extract large subsets of private data at $>$90% accuracy even at the large batch size of 128. Our findings demonstrate that current high-stakes tabular FL is excessively vulnerable to leakage attacks.

****

Low-Variance Gradient Estimation in Unrolled Computation Graphs with ES-Single

Paul Vicol

We propose an evolution strategies-based algorithm for estimating gradients in unrolled computation graphs, called ES-Single. Similarly to the recently-proposed Persistent Evolution Strategies (PES), ES-Single is unbiased, and overcomes chaos arising from recursive function applications by smoothing the meta-loss landscape. ES-Single samples a single perturbation per particle, that is kept fixed over the course of an inner problem (e.g., perturbations are not re-sampled for e

ach partial unroll). Compared to PES, ES-Single is simpler to implement and has lower variance: the variance of ES-Single is constant with respect to the number of truncated unrolls, removing a key barrier in applying ES to long inner problems using short truncations. We show that ES-Single is unbiased for quadratic inner problems, and demonstrate empirically that its variance can be substantially lower than that of PES. ES-Single consistently outperforms PES on a variety of tasks, including a synthetic benchmark task, hyperparameter optimization, training recurrent neural networks, and training learned optimizers.
****
Arithmetic Sampling: Parallel Diverse Decoding for Large Language Models
Luke Vilnis, Yury Zemlyanskiy, Patrick Murray, Alexandre Tachard Passos, Sumit Sanghai
Decoding methods for large language models often trade-off between diversity of outputs and parallelism of computation. Methods such as beam search and Gumbel top-k sampling can guarantee a different output for each element of the beam, but are not easy to parallelize. Alternatively, methods such as temperature sampling and its modifications (top-k sampling, nucleus sampling, typical decoding, and others), are embarrassingly parallel, but have no guarantees about duplicate samples. We present a framework for sampling according to an arithmetic code book implicitly defined by a large language model, compatible with common sampling variations, with provable beam diversity under certain conditions, as well as being embarrassingly parallel and providing unbiased and consistent expectations from the original model. We demonstrate the effectiveness of our approach on WMT machine translation, more than halving the standard deviation when estimating expected BLEU score reward, and closing the BLEU score gap between independent sampling and beam search by up to 63%.
****
Eventual Discounting Temporal Logic Counterfactual Experience Replay
Cameron Voloshin, Abhinav Verma, Yisong Yue
Linear temporal logic (LTL) offers a simplified way of specifying tasks for policy optimization that may otherwise be difficult to describe with scalar reward functions. However, the standard RL framework can be too myopic to find maximally LTL satisfying policies. This paper makes two contributions. First, we develop a new value-function based proxy, using a technique we call eventual discounting, under which one can find policies that satisfy the LTL specification with highest achievable probability. Second, we develop a new experience replay method for generating off-policy data from on-policy rollouts via counterfactual reasoning on different ways of satisfying the LTL specification. Our experiments, conducted in both discrete and continuous state-action spaces, confirm the effectiveness of our counterfactual experience replay approach.
****
Transformers Learn In-Context by Gradient Descent
Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, Joao Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, Max Vladymyrov
At present, the mechanisms of in-context learning in Transformers are not well understood and remain mostly an intuition. In this paper, we suggest that training Transformers on auto-regressive objectives is closely related to gradient-based meta-learning formulations. We start by providing a simple weight construction that shows the equivalence of data transformations induced by 1) a single linear self-attention layer and by 2) gradient-descent (GD) on a regression loss. Motivated by that construction, we show empirically that when training self-attention-only Transformers on simple regression tasks either the models learned by GD and Transformers show great similarity or, remarkably, the weights found by optimization match the construction. Thus we show how trained Transformers become mesa-optimizers i.e. learn models by gradient descent in their forward pass. This allows us, at least in the domain of regression problems, to mechanistically understand the inner workings of in-context learning in optimized Transformers. Building on this insight, we furthermore identify how Transformers surpass the performance of plain gradient descent by learning an iterative curvature correction and learn linear models on deep data representations to solve non-linear regress

ion tasks. Finally, we discuss intriguing parallels to a mechanism identified to be crucial for in-context learning termed induction-head (Olsson et al., 2022) and show how it could be understood as a specific case of in-context learning by gradient descent learning within Transformers.
****

Topological Singularity Detection at Multiple Scales
Julius Von Rohrscheidt, Bastian Rieck
The manifold hypothesis, which assumes that data lies on or close to an unknown manifold of low intrinsic dimension, is a staple of modern machine learning research. However, recent work has shown that real-world data exhibits distinct non-manifold structures, i.e. singularities, that can lead to erroneous findings. Detecting such singularities is therefore crucial as a precursor to interpolation and inference tasks. We address this issue by developing a topological framework that (i) quantifies the local intrinsic dimension, and (ii) yields a Euclidicity score for assessing the 'manifoldness' of a point along multiple scales. Our approach identifies singularities of complex spaces, while also capturing singular structures and local geometric complexity in image data.
****

Improving l1-Certified Robustness via Randomized Smoothing by Leveraging Box Constraints
Vaclav Voracek, Matthias Hein
Randomized smoothing is a popular method to certify robustness of image classifiers to adversarial input perturbations. It is the only certification technique which scales directly to datasets of higher dimension such as ImageNet. However, current techniques are not able to utilize the fact that any adversarial example has to lie in the image space, that is $[0,1]^d$; otherwise, one can trivially detect it. To address this suboptimality, we derive new certification formulae which lead to significant improvements in the certified ℓ_1-robustness without the need of adapting the classifiers or change of smoothing distributions. The code is released at https://github.com/vvoracek/L1-smoothing
****

Vector Quantized Wasserstein Auto-Encoder
Long Tung Vuong, Trung Le, He Zhao, Chuanxia Zheng, Mehrtash Harandi, Jianfei Cai, Dinh Phung
Learning deep discrete latent presentations offers a promise of better symbolic and summarized abstractions that are more useful to subsequent downstream tasks. Inspired by the seminal Vector Quantized Variational Auto-Encoder (VQ-VAE), most of work in learning deep discrete representations has mainly focused on improving the original VQ-VAE form and none of them has studied learning deep discrete representations from the generative viewpoint. In this work, we study learning deep discrete representations from the generative viewpoint. Specifically, we endow discrete distributions over sequences of codewords and learn a deterministic decoder that transports the distribution over the sequences of codewords to the data distribution via minimizing a WS distance between them. We develop further theories to connect it with the clustering viewpoint of WS distance, allowing us to have a better and more controllable clustering solution. Finally, we empirically evaluate our method on several well-known benchmarks, where it achieves better qualitative and quantitative performances than the other VQ-VAE variants in terms of the codebook utilization and image reconstruction/generation.
****

Competitive Gradient Optimization
Abhijeet Vyas, Brian Bullins, Kamyar Azizzadenesheli
We study the problem of convergence to a stationary point in zero-sum games. We propose competitive gradient optimization (CGO), a gradient-based method that incorporates the interactions between two players in zero-sum games for its iterative updates. We provide a continuous-time analysis of CGO and its convergence properties while showing that in the continuous limit, previous methods degenerate to their gradient descent ascent (GDA) variants. We further provide a rate of convergence to stationary points in the discrete-time setting. We propose a generalized class of α-coherent functions and show that for strictly α-

coherent functions, CGO ensures convergence to a saddle point. Moreover, we propose optimistic CGO (oCGO), an optimistic variant, for which we show a convergence rate of $O(\frac{1}{n})$ to saddle points for α-coherent functions.
****

On Provable Copyright Protection for Generative Models
Nikhil Vyas, Sham M. Kakade, Boaz Barak
There is a growing concern that learned conditional generative models may output samples that are substantially similar to some copyrighted data C that was in their training set. We give a formal definition of near access-freeness (NAF) and prove bounds on the probability that a model satisfying this definition outputs a sample similar to C, even if C is included in its training set. Roughly speaking, a generative model p is k-NAF if for every potentially copyrighted data C, the output of p diverges by at most k-bits from the output of a model q that did not access C at all. We also give generative model learning algorithms, which efficiently modify the original generative model learning algorithm in a black box manner, that output generative models with strong bounds on the probability of sampling protected content. Furthermore, we provide promising experiments for both language (transformers) and image (diffusion) generative models, showing minimal degradation in output quality while ensuring strong protections against sampling protected content.
****

Leveraging Offline Data in Online Reinforcement Learning
Andrew Wagenmaker, Aldo Pacchiano
Two central paradigms have emerged in the reinforcement learning (RL) community: online RL and offline RL. In the online RL setting, the agent has no prior knowledge of the environment, and must interact with it in order to find an ϵ-optimal policy. In the offline RL setting, the learner instead has access to a fixed dataset to learn from, but is unable to otherwise interact with the environment, and must obtain the best policy it can from this offline data. Practical scenarios often motivate an intermediate setting: if we have some set of offline data and may also interact with the environment, how can we best use the offline data to minimize the number of online interactions necessary to learn an ϵ-optimal policy. In this work, we consider this setting, which we call the FineTuneRL setting, for MDPs with linear structure. We characterize the necessary number of online samples needed in this setting given access to some offline dataset, and develop an algorithm, FTPedel, which is provably optimal, up to H factors. We show through an explicit example that combining offline data with online interactions can lead to a provable improvement over either purely offline or purely online RL. Finally, our results illustrate the distinction between verifiable learning, the typical setting considered in online RL, and unverifiable learning, the setting often considered in offline RL, and show that there is a formal separation between these regimes.
****

Fast Private Kernel Density Estimation via Locality Sensitive Quantization
Tal Wagner, Yonatan Naamad, Nina Mishra
We study efficient mechanisms for differentially private kernel density estimation (DP-KDE). Prior work for the Gaussian kernel described algorithms that run in time exponential in the number of dimensions d. This paper breaks the exponential barrier, and shows how the KDE can privately be approximated in time linear in d, making it feasible for high-dimensional data. We also present improved bounds for low-dimensional data. Our results are obtained through a general framework, which we term Locality Sensitive Quantization (LSQ), for constructing private KDE mechanisms where existing KDE approximation techniques can be applied. It lets us leverage several efficient non-private KDE methods—like Random Fourier Features, the Fast Gauss Transform, and Locality Sensitive Hashing—and “privatize” them in a black-box manner. Our experiments demonstrate that our resulting DP-KDE mechanisms are fast and accurate on large datasets in both high and low dimensions.
****

Investigating the Role of Model-Based Learning in Exploration and Transfer

Jacob C Walker, Eszter Vértes, Yazhe Li, Gabriel Dulac-Arnold, Ankesh Anand, Theophane Weber, Jessica B Hamrick

State of the art reinforcement learning has enabled training agents on tasks of ever increasing complexity. However, the current paradigm tends to favor training agents from scratch on every new task or on collections of tasks with a view towards generalizing to novel task configurations. The former suffers from poor data efficiency while the latter is difficult when test tasks are out-of-distribution. Agents that can effectively transfer their knowledge about the world pose a potential solution to these issues. In this paper, we investigate transfer learning in the context of model-based agents. Specifically, we aim to understand where exactly environment models have an advantage and why. We find that a model-based approach outperforms controlled model-free baselines for transfer learning. Through ablations, we show that both the policy and dynamics model learnt through exploration matter for successful transfer. We demonstrate our results across three domains which vary in their requirements for transfer: in-distribution procedural (Crafter), in-distribution identical (RoboDesk), and out-of-distribution (Meta-World). Our results show that intrinsic exploration combined with environment models present a viable direction towards agents that are self-supervised and able to generalize to novel reward functions.

****

UPSCALE: Unconstrained Channel Pruning

Alvin Wan, Hanxiang Hao, Kaushik Patnaik, Yueyang Xu, Omer Hadad, David Güera, Zhile Ren, Qi Shan

As neural networks grow in size and complexity, inference speeds decline. To combat this, one of the most effective compression techniques – channel pruning – removes channels from weights. However, for multi-branch segments of a model, channel removal can introduce inference-time memory copies. In turn, these copies increase inference latency – so much so that the pruned model can be slower than the unpruned model. As a workaround, pruners conventionally constrain certain channels to be pruned together. This fully eliminates memory copies but, as we show, significantly impairs accuracy. We now have a dilemma: Remove constraints but increase latency, or add constraints and impair accuracy. In response, our insight is to reorder channels at export time, (1) reducing latency by reducing memory copies and (2) improving accuracy by removing constraints. Using this insight, we design a generic algorithm UPSCALE to prune models with any pruning pattern. By removing constraints from existing pruners, we improve ImageNet accuracy for post-training pruned models by 2.1 points on average – benefiting DenseNet (+16.9), EfficientNetV2 (+7.9), and ResNet (+6.2). Furthermore, by reordering channels, UPSCALE improves inference speeds by up to 2x over a baseline export.

****

Poisoning Language Models During Instruction Tuning

Alexander Wan, Eric Wallace, Sheng Shen, Dan Klein

Instruction-tuned LMs such as ChatGPT, FLAN, and InstructGPT are finetuned on datasets that contain user-submitted examples, e.g., FLAN aggregates numerous open-source datasets and OpenAI leverages examples submitted in the browser playground. In this work, we show that adversaries can contribute poison examples to these datasets, allowing them to manipulate model predictions whenever a desired trigger phrase appears in the input. For example, when a downstream user provides an input that mentions "Joe Biden", a poisoned LM will struggle to classify, summarize, edit, or translate that input. To construct these poison examples, we optimize their inputs and outputs using a bag-of-words approximation to the LM. We evaluate our method on open-source instruction-tuned LMs. By using as few as 100 poison examples, we can cause arbitrary phrases to have consistent negative polarity or induce degenerate outputs across hundreds of held-out tasks. Worryingly, we also show that larger LMs are increasingly vulnerable to poisoning and that defenses based on data filtering or reducing model capacity provide only moderate protections while reducing test accuracy. Notice: This paper contains tasks with obscene content.

****

SeMAIL: Eliminating Distractors in Visual Imitation via Separated Models

Shenghua Wan, Yucen Wang, Minghao Shao, Ruying Chen, De-Chuan Zhan
Model-based imitation learning (MBIL) is a popular reinforcement learning method that improves sample efficiency on high-dimension input sources, such as images and videos. Following the convention of MBIL research, existing algorithms are highly deceptive by task-irrelevant information, especially moving distractors in videos. To tackle this problem, we propose a new algorithm - named Separated Model-based Adversarial Imitation Learning (SeMAIL) - decoupling the environment dynamics into two parts by task-relevant dependency, which is determined by agent actions, and training separately. In this way, the agent can imagine its trajectories and imitate the expert behavior efficiently in task-relevant state space. Our method achieves near-expert performance on various visual control tasks with complex observations and the more challenging tasks with different backgrounds from expert observations.
****

Multiplier Bootstrap-based Exploration
Runzhe Wan, Haoyu Wei, Branislav Kveton, Rui Song
Despite the great interest in the bandit problem, designing efficient algorithms for complex models remains challenging, as there is typically no analytical way to quantify uncertainty. In this paper, we propose Multiplier Bootstrap-based Exploration (MBE), a novel exploration strategy that is applicable to any reward model amenable to weighted loss minimization. We prove both instance-dependent and instance-independent rate-optimal regret bounds for MBE in sub-Gaussian multi-armed bandits. With extensive simulation and real-data experiments, we show the generality and adaptivity of MBE.
****

Bandit Multi-linear DR-Submodular Maximization and Its Applications on Adversarial Submodular Bandits
Zongqi Wan, Jialin Zhang, Wei Chen, Xiaoming Sun, Zhijie Zhang
We investigate the online bandit learning of the monotone multi-linear DR-submodular functions, designing the algorithm $\mathtt{BanditMLSM}$ that attains $O(T^{2/3}\log T)$ of $(1-1/e)$-regret. Then we reduce submodular bandit with partition matroid constraint and bandit sequential monotone maximization to the online bandit learning of the monotone multi-linear DR-submodular functions, attaining $O(T^{2/3}\log T)$ of $(1-1/e)$-regret in both problems, which improve the existing results. To the best of our knowledge, we are the first to give a sublinear regret algorithm for the submodular bandit with partition matroid constraint. A special case of this problem is studied by Streeter et al.(2009). They prove a $O(T^{4/5})$ $(1-1/e)$-regret upper bound. For the bandit sequential submodular maximization, the existing work proves an $O(T^{2/3})$ regret with a suboptimal $1/2$ approximation ratio (Niazadeh et al. 2021).
****

Tight Regret Bounds for Single-pass Streaming Multi-armed Bandits
Chen Wang
Regret minimization in streaming multi-armed bandits (MABs) has been studied extensively, and recent work has shown that algorithms with $o(K)$ memory have to incur $\Omega(T^{2/3})$ regret, where K and T are the numbers of arms and trials. However, the previous best regret upper bound is still $O(K^{1/3} T^{2/3}\log^{1/3}(T))$, which is achieved by the simple uniform exploration algorithm. In this paper, we close this gap and complete the picture of regret minimization in single-pass streaming MABs. We first improve the regret lower bound to $\Omega(K^{1/3}T^{2/3})$ for algorithms with $o(K)$ memory. We then show that the $\log^{1/3}(T)$ factor is not necessary by designing algorithms with at most $O(\log^*(K))$-arm memory and achieve $O(K^{1/3}T^{2/3})$ expected regret based on streaming ε-best arm algorithms. We further tested the empirical performances of our algorithms on simulated MABs instances, where the proposed algorithms outperform the benchmark uniform exploration algorithm by a large margin and, on occasion, reduce the regret by up to 70%.
****

Improved Active Multi-Task Representation Learning via Lasso
Yiping Wang, Yifang Chen, Kevin Jamieson, Simon Shaolei Du

To leverage the copious amount of data from source tasks and overcome the scarcity of the target task samples, representation learning based on multi-task pretraining has become a standard approach in many applications. However, up until now, most existing works design a source task selection strategy from a purely empirical perspective. Recently, Chen et al., 2022 gave the first active multi-task representation learning (A-MTRL) algorithm which adaptively samples from source tasks and can provably reduce the total sample complexity using the L2-regularized-target-source-relevance parameter ν^2. But their work is theoretically suboptimal in terms of total source sample complexity and is less practical in some real-world scenarios where sparse training source task selection is desired. In this paper, we address both issues. Specifically, we show the strict dominance of the L1-regularized-relevance-based (ν^1-based) strategy by giving a lower bound for the ν^2-based strategy. When ν^1 is unknown, we propose a practical algorithm that uses the LASSO program to estimate ν^1. Our algorithm successfully recovers the optimal result in the known case. In addition to our sample complexity results, we also characterize the potential of our ν^1-based strategy in sample-cost-sensitive settings. Finally, we provide experiments on real-world computer vision datasets to illustrate the effectiveness of our proposed method.

****

Tilted Sparse Additive Models

Yingjie Wang, Hong Chen, Weifeng Liu, Fengxiang He, Tieliang Gong, Youcheng Fu, Dacheng Tao

Additive models have been burgeoning in data analysis due to their flexible representation and desirable interpretability. However, most existing approaches are constructed under empirical risk minimization (ERM), and thus perform poorly in situations where average performance is not a suitable criterion for the problems of interest, e.g., data with complex non-Gaussian noise, imbalanced labels or both of them. In this paper, a novel class of sparse additive models is proposed under tilted empirical risk minimization (TERM), which addresses the deficiencies in ERM by imposing tilted impact on individual losses, and is flexibly capable of achieving a variety of learning objectives, e.g., variable selection, robust estimation, imbalanced classification and multiobjective learning. On the theoretical side, a learning theory analysis which is centered around the generalization bound and function approximation error bound (under some specific data distributions) is conducted rigorously. On the practical side, an accelerated optimization algorithm is designed by integrating Prox-SVRG and random Fourier acceleration technique. The empirical assessments verify the competitive performance of our approach on both synthetic and real data.

****

From Hypergraph Energy Functions to Hypergraph Neural Networks

Yuxin Wang, Quan Gan, Xipeng Qiu, Xuanjing Huang, David Wipf

Hypergraphs are a powerful abstraction for representing higher-order interactions between entities of interest. To exploit these relationships in making downstream predictions, a variety of hypergraph neural network architectures have recently been proposed, in large part building upon precursors from the more traditional graph neural network (GNN) literature. Somewhat differently, in this paper we begin by presenting an expressive family of parameterized, hypergraph-regularized energy functions. We then demonstrate how minimizers of these energies effectively serve as node embeddings that, when paired with a parameterized classifier, can be trained end-to-end via a supervised bilevel optimization process. Later, we draw parallels between the implicit architecture of the predictive models emerging from the proposed bilevel hypergraph optimization, and existing GNN architectures in common use. Empirically, we demonstrate state-of-the-art results on various hypergraph node classification benchmarks. Code is available at https://github.com/yxzwang/PhenomNN.

****

A Closer Look at Self-Supervised Lightweight Vision Transformers

Shaoru Wang, Jin Gao, Zeming Li, Xiaoqin Zhang, Weiming Hu

Self-supervised learning on large-scale Vision Transformers (ViTs) as pre-traini

ng methods has achieved promising downstream performance. Yet, how much these pre-training paradigms promote lightweight ViTs' performance is considerably less studied. In this work, we develop and benchmark several self-supervised pre-training methods on image classification tasks and some downstream dense prediction tasks. We surprisingly find that if proper pre-training is adopted, even vanilla lightweight ViTs show comparable performance to previous SOTA networks with delicate architecture design. It breaks the recently popular conception that vanilla ViTs are not suitable for vision tasks in lightweight regimes. We also point out some defects of such pre-training, e.g., failing to benefit from large-scale pre-training data and showing inferior performance on data-insufficient downstream tasks. Furthermore, we analyze and clearly show the effect of such pre-training by analyzing the properties of the layer representation and attention maps for related models. Finally, based on the above analyses, a distillation strategy during pre-training is developed, which leads to further downstream performance improvement for MAE-based pre-training. Code is available at https://github.com/wangsr126/mae-lite.

****

PreNAS: Preferred One-Shot Learning Towards Efficient Neural Architecture Search

Haibin Wang, Ce Ge, Hesen Chen, Xiuyu Sun

The wide application of pre-trained models is driving the trend of once-for-all training in one-shot neural architecture search (NAS). However, training within a huge sample space damages the performance of individual subnets and requires much computation to search for a optimal model. In this paper, we present PreNAS, a search-free NAS approach that accentuates target models in one-shot training. Specifically, the sample space is dramatically reduced in advance by a zero-cost selector, and weight-sharing one-shot training is performed on the preferred architectures to alleviate update conflicts. Extensive experiments have demonstrated that PreNAS consistently outperforms state-of-the-art one-shot NAS competitors for both Vision Transformer and convolutional architectures, and importantly, enables instant specialization with zero search cost. Our code is available at https://github.com/tinyvision/PreNAS.

****

Adversarial Policies Beat Superhuman Go AIs

Tony Tong Wang, Adam Gleave, Tom Tseng, Kellin Pelrine, Nora Belrose, Joseph Miller, Michael D Dennis, Yawen Duan, Viktor Pogrebniak, Sergey Levine, Stuart Russell

We attack the state-of-the-art Go-playing AI system KataGo by training adversarial policies against it, achieving a $>$97% win rate against KataGo running at superhuman settings. Our adversaries do not win by playing Go well. Instead, they trick KataGo into making serious blunders. Our attack transfers zero-shot to other superhuman Go-playing AIs, and is comprehensible to the extent that human experts can implement it without algorithmic assistance to consistently beat superhuman AIs. The core vulnerability uncovered by our attack persists even in KataGo agents adversarially trained to defend against our attack. Our results demonstrate that even superhuman AI systems may harbor surprising failure modes. Example games are available https://goattack.far.ai/.

****

On Regularization and Inference with Label Constraints

Kaifu Wang, Hangfeng He, Tin D. Nguyen, Piyush Kumar, Dan Roth

Prior knowledge and symbolic rules in machine learning are often expressed in the form of label constraints, especially in structured prediction problems. In this work, we compare two common strategies for encoding label constraints in a machine learning pipeline, regularization with constraints and constrained inference, by quantifying their impact on model performance. For regularization, we show that it narrows the generalization gap by precluding models that are inconsistent with the constraints. However, its preference for small violations introduces a bias toward a suboptimal model. For constrained inference, we show that it reduces the population risk by correcting a model's violation, and hence turns the violation into an advantage. Given these differences, we further explore the use of two approaches together and propose conditions for constrained inference t

o compensate for the bias introduced by regularization, aiming to improve both t
he model complexity and optimal risk.
****

Policy Gradient in Robust MDPs with Global Convergence Guarantee
Qiuhao Wang, Chin Pang Ho, Marek Petrik
Robust Markov decision processes (RMDPs) provide a promising framework for compu
ting reliable policies in the face of model errors. Many successful reinforcemen
t learning algorithms build on variations of policy-gradient methods, but adapti
ng these methods to RMDPs has been challenging. As a result, the applicability o
f RMDPs to large, practical domains remains limited. This paper proposes a new D
ouble-Loop Robust Policy Gradient (DRPG), the first generic policy gradient meth
od for RMDPs. In contrast with prior robust policy gradient algorithms, DRPG mon
otonically reduces approximation errors to guarantee convergence to a globally o
ptimal policy in tabular RMDPs. We introduce a novel parametric transition kerne
l and solve the inner loop robust policy via a gradient-based method. Finally, o
ur numerical results demonstrate the utility of our new algorithm and confirm it
s global convergence properties.
****

Adaptive Smoothing Gradient Learning for Spiking Neural Networks
Ziming Wang, Runhao Jiang, Shuang Lian, Rui Yan, Huajin Tang
Spiking neural networks (SNNs) with biologically inspired spatio-temporal dynami
cs demonstrate superior energy efficiency on neuromorphic architectures. Error b
ackpropagation in SNNs is prohibited by the all-or-none nature of spikes. The ex
isting solution circumvents this problem by a relaxation on the gradient calcula
tion using a continuous function with a constant relaxation de- gree, so-called
surrogate gradient learning. Nevertheless, such a solution introduces additional
 smoothing error on spike firing which leads to the gradients being estimated in
accurately. Thus, how to adaptively adjust the relaxation degree and eliminate s
moothing error progressively is crucial. Here, we propose a methodology such tha
t training a prototype neural network will evolve into training an SNN gradually
 by fusing the learnable relaxation degree into the network with random spike no
ise. In this way, the network learns adaptively the accurate gradients of loss l
andscape in SNNs. The theoretical analysis further shows optimization on such a
noisy network could be evolved into optimization on the embedded SNN with shared
 weights progressively. Moreover, The experiments on static images, dynamic even
t streams, speech, and instrumental sounds show the proposed method achieves sta
te-of-the-art performance across all the datasets with remarkable robustness on
different relaxation degrees.
****

CircuitNet: A Generic Neural Network to Realize Universal Circuit Motif Modeling
Yansen Wang, Xinyang Jiang, Kan Ren, Caihua Shan, Xufang Luo, Dongqi Han, Kaitao
 Song, Yifei Shen, Dongsheng Li
The successes of artificial neural networks (ANNs) are largely attributed to mim
icking the human brain structures. Recent advances in neuroscience revealed that
 neurons interact with each other through various kinds of connectivity patterns
 to process information, in which the common connectivity patterns are also call
ed circuit motifs. However, many existing ANNs can only model one or two circuit
 motifs in their architectures, so that their performance may drastically vary a
mong different types of machine learning tasks. In this paper, we propose a new
type of neural network inspired by the architectures of neuronal circuits, namel
y Circuit Neural Network (CircuitNet). In CircuitNet, a group of densely connect
ed neurons, namely circuit motif unit (CMU), form the basic unit of the network,
 which is capable of modeling universal circuit motifs by adjusting the weights
within the CMUs. Compared with traditional feed-forward networks, CircuitNet has
 the ability to model more types of neuron connections such as feed-back and lat
eral motifs. Inspired by the locally dense and globally sparse structure of the
human brain, several iterations of signal transmission among different CMUs are
achieved by sparse connections through the input ports and output ports of diffe
rent CMUs. Experiments have demonstrated that CircuitNet can outperform popular
neural network architectures in function approximation, reinforcement learning,

image classification, and time series forecasting tasks.
****
Generalized Polyak Step Size for First Order Optimization with Momentum
Xiaoyu Wang, Mikael Johansson, Tong Zhang
In machine learning applications, it is well known that carefully designed learning rate (step size) schedules can significantly improve the convergence of commonly used first-order optimization algorithms. Therefore how to set step size adaptively becomes an important research question. A popular and effective method is the Polyak step size, which sets step size adaptively for gradient descent or stochastic gradient descent without the need to estimate the smoothness parameter of the objective function. However, there has not been a principled way to generalize the Polyak step size for algorithms with momentum accelerations. This paper presents a general framework to set the learning rate adaptively for first-order optimization methods with momentum, motivated by the derivation of Polyak step size. It is shown that the resulting techniques are much less sensitive to the choice of momentum parameter and may avoid the oscillation of the heavy-ball method on ill-conditioned problems. These adaptive step sizes are further extended to the stochastic settings, which are attractive choices for stochastic gradient descent with momentum. Our methods are demonstrated to be more effective for stochastic gradient methods than prior adaptive step size algorithms in large-scale machine learning tasks.
****
Near-Minimax-Optimal Risk-Sensitive Reinforcement Learning with CVaR
Kaiwen Wang, Nathan Kallus, Wen Sun
In this paper, we study risk-sensitive Reinforcement Learning (RL), focusing on the objective of Conditional Value at Risk (CVaR) with risk tolerance τ. Starting with multi-arm bandits (MABs), we show the minimax CVaR regret rate is $\Omega(\sqrt{\tau^{-1}AK})$, where A is the number of actions and K is the number of episodes, and that it is achieved by an Upper Confidence Bound algorithm with a novel Bernstein bonus. For online RL in tabular Markov Decision Processes (MDPs), we show a minimax regret lower bound of $\Omega(\sqrt{\tau^{-1}SAK})$ (with normalized cumulative rewards), where S is the number of states, and we propose a novel bonus-driven Value Iteration procedure. We show that our algorithm achieves the optimal regret of $\widetilde O(\sqrt{\tau^{-1}SAK})$ under a continuity assumption and in general attains a near-optimal regret of $\widetilde O(\tau^{-1}\sqrt{SAK})$, which is minimax-optimal for constant τ. This improves on the best available bounds. By discretizing rewards appropriately, our algorithms are computationally efficient.
****
FedHPO-Bench: A Benchmark Suite for Federated Hyperparameter Optimization
Zhen Wang, Weirui Kuang, Ce Zhang, Bolin Ding, Yaliang Li
Research in the field of hyperparameter optimization (HPO) has been greatly accelerated by existing HPO benchmarks. Nonetheless, existing efforts in benchmarking all focus on HPO for traditional learning paradigms while ignoring federated learning (FL), a promising paradigm for collaboratively learning models from dispersed data. In this paper, we first identify some uniqueness of federated hyperparameter optimization (FedHPO) from various aspects, showing that existing HPO benchmarks no longer satisfy the need to study FedHPO methods. To facilitate the research of FedHPO, we propose and implement a benchmark suite FedHPO-Bench that incorporates comprehensive FedHPO problems, enables flexible customization of the function evaluations, and eases continuing extensions. We conduct extensive experiments based on FedHPO-Bench to provide the community with more insights into FedHPO. We open-sourced FedHPO-Bench at https://github.com/alibaba/FederatedScope/tree/master/benchmark/FedHPOBench.
****
A/B Testing in Network Data with Covariate-Adaptive Randomization
Jialu Wang, Ping Li, Feifang Hu
Users linked together through a network often tend to have similar behaviors. This phenomenon is usually known as network interaction. Users’ characteristics, the covariates, are often correlated with their outcomes. Therefore, one should i

ncorporate both the covariates and the network information in a carefully design ed randomization to improve the estimation of the average treatment effect (ATE) in network A/B testing. In this paper, we propose a new adaptive procedure to b alance both the network and the covariates. We show that the imbalance measures with respect to the covariates and the network are $O_p(1)$. We also demonstrate the relationships between the improved balances and the increased efficiency in terms of the mean square error (MSE). Numerical studies demonstrate the advance d performance of the proposed procedure regarding the greater comparability of t he treatment groups and the reduction of MSE for estimating the ATE.
****

Learning Belief Representations for Partially Observable Deep RL
Andrew Wang, Andrew C Li, Toryn Q. Klassen, Rodrigo Toro Icarte, Sheila A. Mcilr aith
Many important real-world Reinforcement Learning (RL) problems involve partial o bservability and require policies with memory. Unfortunately, standard deep RL a lgorithms for partially observable settings typically condition on the full hist ory of interactions and are notoriously difficult to train. We propose a novel d eep, partially observable RL algorithm based on modelling belief states — a tech nique typically used when solving tabular POMDPs, but that has traditionally bee n difficult to apply to more complex environments. Our approach simplifies polic y learning by leveraging state information at training time, that may not be ava ilable at deployment time. We do so in two ways: first, we decouple belief state modelling (via unsupervised learning) from policy optimization (via RL); and se cond, we propose a representation learning approach to capture a compact set of reward-relevant features of the state. Experiments demonstrate the efficacy of o ur approach on partially observable domains requiring information seeking and lo ng-term memory.
****

Warm-Start Actor-Critic: From Approximation Error to Sub-optimality Gap
Hang Wang, Sen Lin, Junshan Zhang
Warm-Start reinforcement learning (RL), aided by a prior policy obtained from of fline training, is emerging as a promising RL approach for practical application s. Recent empirical studies have demonstrated that the performance of Warm-Start RL can be improved quickly in some cases but become stagnant in other cases, es pecially when the function approximation is used. To this end, the primary objec tive of this work is to build a fundamental understanding on ”whether and when o nline learning can be significantly accelerated by a warm-start policy from offl ine RL?”. Specifically, we consider the widely used Actor-Critic (A-C) method wi th a prior policy. We first quantify the approximation errors in the Actor updat e and the Critic update, respectively. Next, we cast the Warm-Start A-C algorith m as Newton’s method with perturbation, and study the impact of the approximatio n errors on the finite-time learning performance with inaccurate Actor/Critic up dates. Under some general technical conditions, we derive the upper bounds, whic h shed light on achieving the desired finite-learning performance in the Warm-St art A-C algorithm. In particular, our findings reveal that it is essential to re duce the algorithm bias in online learning. We also obtain lower bounds on the s ub-optimality gap of the Warm-Start A-C algorithm to quantify the impact of the bias and error propagation.
****

Slot-VAE: Object-Centric Scene Generation with Slot Attention
Yanbo Wang, Letao Liu, Justin Dauwels
Slot attention has shown remarkable object-centric representation learning perfo rmance in computer vision tasks without requiring any supervision. Despite its o bject-centric binding ability brought by compositional modelling, as a determini stic module, slot attention lacks the ability to generate novel scenes. In this paper, we propose the Slot-VAE, a generative model that integrates slot attentio n with the hierarchical VAE framework for object-centric structured scene genera tion. For each image, the model simultaneously infers a global scene representat ion to capture high-level scene structure and object-centric slot representation s to embed individual object components. During generation, slot representations

are generated from the global scene representation to ensure coherent scene structures. Our extensive evaluation of the scene generation ability indicates that Slot-VAE outperforms slot representation-based generative baselines in terms of sample quality and scene structure accuracy.

****

DIVISION: Memory Efficient Training via Dual Activation Precision

Guanchu Wang, Zirui Liu, Zhimeng Jiang, Ninghao Liu, Na Zou, Xia Hu

Activation compressed training provides a solution towards reducing the memory cost of training deep neural networks (DNNs). However, state-of-the-art work combines a search of quantization bit-width with the training, which makes the procedure complicated and less transparent. To this end, we propose a simple and effective method to compress DNN training. Our method is motivated by an instructive observation: DNN backward propagation mainly utilizes the low-frequency component (LFC) of the activation maps, while the majority of memory is for caching the high-frequency component (HFC) during the training. This indicates the HFC of activation maps is highly redundant and compressible, which inspires our proposed Dual Activation Precision (DIVISION). During the training, DIVISION preserves a high-precision copy of LFC and compresses the HFC into a light-weight copy with low numerical precision. This can significantly reduce the memory cost while maintaining a competitive model accuracy. Experiment results show DIVISION has better comprehensive performance than state-of-the-art methods, including over 10x compression of activation maps and competitive training throughput, without loss of model accuracy. The source code is available at https://github.com/guanchuwang/division.

****

CocktailSGD: Fine-tuning Foundation Models over 500Mbps Networks

Jue Wang, Yucheng Lu, Binhang Yuan, Beidi Chen, Percy Liang, Christopher De Sa, Christopher Re, Ce Zhang

Distributed training of foundation models, especially large language models (LLMs), is communication-intensive and so has heavily relied on centralized data centers with fast interconnects. Can we train on slow networks and unlock the potential of decentralized infrastructure for foundation models? In this paper, we propose CocktailSGD, a novel communication-efficient training framework that combines three distinct compression techniques – random sparsification, top-K sparsification, and quantization – to achieve much greater compression than each individual technique alone. We justify the benefit of such a hybrid approach through a theoretical analysis of convergence. Empirically, we show that CocktailSGD achieves up to 117$\times$ compression in fine-tuning LLMs up to 20 billion parameters without hurting convergence. On a 500Mbps network, CocktailSGD only incurs $\sim$1.2$\times$ slowdown compared with data center networks.

****

Magneto: A Foundation Transformer

Hongyu Wang, Shuming Ma, Shaohan Huang, Li Dong, Wenhui Wang, Zhiliang Peng, Yu Wu, Payal Bajaj, Saksham Singhal, Alon Benhaim, Barun Patra, Zhun Liu, Vishrav Chaudhary, Xia Song, Furu Wei

A big convergence of model architectures across language, vision, speech, and multimodal is emerging. However, under the same name ”Transformers”, the above areas use different implementations for better performance, e.g., Post-LayerNorm for BERT, and Pre-LayerNorm for GPT and vision Transformers. We call for the development of Foundation Transformer for true general-purpose modeling, which serves as a go-to architecture for various tasks and modalities with guaranteed training stability. In this work, we introduce a Transformer variant, named Magneto, to fulfill the goal. Specifically, we propose Sub-LayerNorm for good expressivity, and the initialization strategy theoretically derived from DeepNet for stable scaling up. Extensive experiments demonstrate its superior performance and better stability than the de facto Transformer variants designed for various applications, including language modeling (i.e., BERT, and GPT), machine translation, vision pretraining (i.e., BEiT), speech recognition, and multimodal pretraining (i.e., BEiT-3).

****

Direct Parameterization of Lipschitz-Bounded Deep Networks
Ruigang Wang, Ian Manchester
This paper introduces a new parameterization of deep neural networks (both fully-connected and convolutional) with guaranteed ℓ^2 Lipschitz bounds, i.e. limited sensitivity to input perturbations. The Lipschitz guarantees are equivalent to the tightest-known bounds based on certification via a semidefinite program (SDP). We provide a “direct” parameterization, i.e., a smooth mapping from $\mathbb R^N$ onto the set of weights satisfying the SDP-based bound. Moreover, our parameterization is complete, i.e. a neural network satisfies the SDP bound if and only if it can be represented via our parameterization. This enables training using standard gradient methods, without any inner approximation or computationally intensive tasks (e.g. projections or barrier terms) for the SDP constraint. The new parameterization can equivalently be thought of as either a new layer type (the sandwich layer), or a novel parameterization of standard feedforward networks with parameter sharing between neighbouring layers. A comprehensive set of experiments on image classification shows that sandwich layers outperform previous approaches on both empirical and certified robust accuracy. Code is available at https://github.com/acfr/LBDN.
****
Tighter Information-Theoretic Generalization Bounds from Supersamples
Ziqiao Wang, Yongyi Mao
In this work, we present a variety of novel information-theoretic generalization bounds for learning algorithms, from the supersample setting of Steinke & Zakynthinou (2020)—the setting of the “conditional mutual information” framework. Our development exploits projecting the loss pair (obtained from a training instance and a testing instance) down to a single number and correlating loss values with a Rademacher sequence (and its shifted variants). The presented bounds include square-root bounds, fast-rate bounds, including those based on variance and sharpness, and bounds for interpolating algorithms etc. We show theoretically or empirically that these bounds are tighter than all information-theoretic bounds known to date on the same supersample setting.
****
NP-SemiSeg: When Neural Processes meet Semi-Supervised Semantic Segmentation
Jianfeng Wang, Daniela Massiceti, Xiaolin Hu, Vladimir Pavlovic, Thomas Lukasiewicz
Semi-supervised semantic segmentation involves assigning pixel-wise labels to unlabeled images at training time. This is useful in a wide range of real-world applications where collecting pixel-wise labels is not feasible in time or cost. Current approaches to semi-supervised semantic segmentation work by predicting pseudo-labels for each pixel from a class-wise probability distribution output by a model. If this predicted probability distribution is incorrect, however, it leads to poor segmentation results which can have knock-on consequences in safety critical systems, like medical images or self-driving cars. It is, therefore, important to understand what a model does not know, which is mainly achieved by uncertainty quantification. Recently, neural processes (NPs) have been explored in semi-supervised image classification, and they have been a computationally efficient and effective method for uncertainty quantification. In this work, we move one step forward by adapting NPs to semi-supervised semantic segmentation, resulting in a new model called NP-SemiSeg. We experimentally evaluated NP-SemiSeg on the public benchmarks PASCAL VOC 2012 and Cityscapes, with different training settings, and the results verify its effectiveness.
****
GC-Flow: A Graph-Based Flow Network for Effective Clustering
Tianchun Wang, Farzaneh Mirzazadeh, Xiang Zhang, Jie Chen
Graph convolutional networks (GCNs) are discriminative models that directly model the class posterior $p(y|\mathbf{x})$ for semi-supervised classification of graph data. While being effective, as a representation learning approach, the node representations extracted from a GCN often miss useful information for effective clustering, because the objectives are different. In this work, we design normalizing flows that replace GCN layers, leading to a generative model that models

both the class conditional likelihood $p(\mathbf{x}|y)$ and the class prior $p(y)$. The resulting neural network, GC-Flow, retains the graph convolution operations while being equipped with a Gaussian mixture representation space. It enjoys two benefits: it not only maintains the predictive power of GCN, but also produces well-separated clusters, due to the structuring of the representation space. We demonstrate these benefits on a variety of benchmark data sets. Moreover, we show that additional parameterization, such as that on the adjacency matrix used for graph convolutions, yields additional improvement in clustering.
****

Curriculum Co-disentangled Representation Learning across Multiple Environments for Social Recommendation

Xin Wang, Zirui Pan, Yuwei Zhou, Hong Chen, Chendi Ge, Wenwu Zhu

There exist complex patterns behind the decision-making processes of different individuals across different environments. For instance, in a social recommender system, various user behaviors are driven by highly entangled latent factors from two environments, i.e., consuming environment where users consume items and social environment where users connect with each other. Uncovering the disentanglement of these latent factors for users can benefit in enhanced explainability and controllability for recommendation. However, in literature there has been no work on social recommendation capable of disentangling user representations across consuming and social environments. To solve this problem, we study co-disentangled representation learning across different environments via proposing the curriculum co-disentangled representation learning (CurCoDis) model to disentangle the hidden factors for users across both consuming and social environments. To co-disentangle joint representations for user-item consumption and user-user social graph simultaneously, we partition the social graph into equal-size sub-graphs with minimum number of edges being cut, and design a curriculum weighing strategy for subgraph training through measuring the complexity of subgraphs via Descartes’ rule of signs. We further develop the prototype-routing optimization mechanism, which achieves co-disentanglement of user representations across consuming and social environments. Extensive experiments for social recommendation demonstrate that our proposed CurCoDis model can significantly outperform state-of-the-art methods on several real-world datasets.
****

Data Efficient Neural Scaling Law via Model Reusing

Peihao Wang, Rameswar Panda, Zhangyang Wang

The number of parameters in large transformers has been observed to grow exponentially. Despite notable performance improvements, concerns have been raised that such a growing model size will run out of data in the near future. As manifested in the neural scaling law, modern learning backbones are not data-efficient. To maintain the utility of the model capacity, training data should be increased proportionally. In this paper, we study the neural scaling law under the previously overlooked data scarcity regime, focusing on the more challenging situation where we need to train a gigantic model with a disproportionately limited supply of available training data. We find that the existing power laws underestimate the data inefficiency of large transformers. Their performance will drop significantly if the training set is insufficient. Fortunately, we discover another blessing - such a data-inefficient scaling law can be restored through a model reusing approach that warm-starts the training of a large model by initializing it using smaller models. Our empirical study shows that model reusing can effectively reproduce the power law under the data scarcity regime. When progressively applying model reusing to expand the model size, we also observe consistent performance improvement in large transformers. We release our code at: https://github.com/VITA-Group/Data-Efficient-Scaling.
****

Deep Temporal Sets with Evidential Reinforced Attentions for Unique Behavioral Pattern Discovery

Dingrong Wang, Deep Shankar Pandey, Krishna Prasad Neupane, Zhiwei Yu, Ervine Zheng, Zhi Zheng, Qi Yu

Machine learning-driven human behavior analysis is gaining attention in behavior

al/mental healthcare, due to its potential to identify behavioral patterns that cannot be recognized by traditional assessments. Real-life applications, such as digital behavioral biomarker identification, often require the discovery of complex spatiotemporal patterns in multimodal data, which is largely under-explored. To fill this gap, we propose a novel model that integrates uniquely designed Deep Temporal Sets (DTS) with Evidential Reinforced Attentions (ERA). DTS captures complex temporal relationships in the input and generates a set-based representation, while ERA captures the policy network’s uncertainty and conducts evidence-aware exploration to locate attentive regions in behavioral data. Using child-computer interaction data as a testing platform, we demonstrate the effectiveness of DTS-ERA in differentiating children with Autism Spectrum Disorder and typically developing children based on sequential multimodal visual and touch behaviors. Comparisons with baseline methods show that our model achieves superior performance and has the potential to provide objective, quantitative, and precise analysis of complex human behaviors.
****
Active Learning based Structural Inference
Aoran Wang, Jun Pang
In this paper, we propose a novel framework, Active Learning based Structural Inference (ALaSI), to infer the existence of directed connections from observed agents’ states over a time period in a dynamical system. With the help of deep active learning, ALaSI is competent in learning the representation of connections with a relatively small pool of prior knowledge. Moreover, based on information theory, the proposed inter- and out-of-scope message learning pipelines are remarkably beneficial to structural inference for large dynamical systems. We evaluate ALaSI on various large datasets including simulated systems and real-world networks, to demonstrate that ALaSI is able to outperform previous methods in precisely inferring the existence of connections in large systems under either supervised learning or unsupervised learning.
****
Better Diffusion Models Further Improve Adversarial Training
Zekai Wang, Tianyu Pang, Chao Du, Min Lin, Weiwei Liu, Shuicheng Yan
It has been recognized that the data generated by the denoising diffusion probabilistic model (DDPM) improves adversarial training. After two years of rapid development in diffusion models, a question naturally arises: can better diffusion models further improve adversarial training? This paper gives an affirmative answer by employing the most recent diffusion model which has higher efficiency (~ 20 sampling steps) and image quality (lower FID score) compared with DDPM. Our adversarially trained models achieve state-of-the-art performance on RobustBench using only generated data (no external datasets). Under the ℓ_∞-norm threat model with $\epsilon=8/255$, our models achieve $70.69\\%$ and $42.67\\%$ robust accuracy on CIFAR-10 and CIFAR-100, respectively, i.e. improving upon previous state-of-the-art models by $+4.58\\%$ and $+8.03\\%$. Under the ℓ_2-norm threat model with $\epsilon=128/255$, our models achieve $84.86\\%$ on CIFAR-10 ($+4.44\\%$). These results also beat previous works that use external data. We also provide compelling results on the SVHN and TinyImageNet datasets. Our code is at https://github.com/wzekai99/DM-Improves-AT.
****
Polarity Is All You Need to Learn and Transfer Faster
Qingyang Wang, Michael Alan Powell, Eric W Bridgeford, Ali Geisa, Joshua T Vogelstein
Natural intelligences (NIs) thrive in a dynamic world - they learn quickly, sometimes with only a few samples. In contrast, artificial intelligences (AIs) typically learn with a prohibitive number of training samples and computational power. What design principle difference between NI and AI could contribute to such a discrepancy? Here, we investigate the role of weight polarity: development processes initialize NIs with advantageous polarity configurations; as NIs grow and learn, synapse magnitudes update, yet polarities are largely kept unchanged. We demonstrate with simulation and image classification tasks that if weight polarities are adequately set a priori, then networks learn with less time and data. We

also explicitly illustrate situations in which a priori setting the weight polarities is disadvantageous for networks. Our work illustrates the value of weight polarities from the perspective of statistical and computational efficiency during learning.

****

Projected Tensor Power Method for Hypergraph Community Recovery

Jinxin Wang, Yuen-Man Pun, Xiaolu Wang, Peng Wang, Anthony Man-Cho So

This paper investigates the problem of exact community recovery in the symmetric d-uniform $(d \geq 2)$ hypergraph stochastic block model (d-HSBM). In this model, a d-uniform hypergraph with n nodes is generated by first partitioning the n nodes into $K\geq 2$ equal-sized disjoint communities and then generating hyperedges with a probability that depends on the community memberships of d nodes. Despite the non-convex and discrete nature of the maximum likelihood estimation problem, we develop a simple yet efficient iterative method, called the projected tensor power method, to tackle it. As long as the initialization satisfies a partial recovery condition in the logarithmic degree regime of the problem, we show that our proposed method can exactly recover the hidden community structure down to the information-theoretic limit with high probability. Moreover, our proposed method exhibits a competitive time complexity of $\mathcal{O}(n\log^2n/\log\log n)$ when the aforementioned initialization condition is met. We also conduct numerical experiments to validate our theoretical findings.

****

Estimating Possible Causal Effects with Latent Variables via Adjustment

Tian-Zuo Wang, Tian Qin, Zhi-Hua Zhou

Causal effect identification is a fundamental task in artificial intelligence. A most ideal scenario for causal effect identification is that there is a directed acyclic graph as a prior causal graph encoding the causal relations of all relevant variables. In real tasks, however, the prior causal graph is usually not available, and some relevant variables may be latent as well. With observational data, we can only learn a partial ancestral graph (PAG), which contains some indeterminate causal relations. Since many causal graphs can correspond to one PAG, they are possibly associated with different causal effects. The aim of this paper is to estimate these possible causal effects via covariate adjustment given a PAG. This task is challenging because the number of causal graphs corresponding to a PAG grows super-exponentially with the number of variables. We propose a new graphical characterization for possible adjustment sets, and based on this, we develop the first method to determine the set of possible causal effects that are consistent with the given PAG without enumerating any causal graphs. Our method can output the same set as the enumeration method with super-exponentially less complexity. Experiments validate the effectiveness and tremendous efficiency improvement of the proposed method.

****

InfoDiffusion: Representation Learning Using Information Maximizing Diffusion Models

Yingheng Wang, Yair Schiff, Aaron Gokaslan, Weishen Pan, Fei Wang, Christopher De Sa, Volodymyr Kuleshov

While diffusion models excel at generating high-quality samples, their latent variables typically lack semantic meaning and are not suitable for representation learning. Here, we propose InfoDiffusion, an algorithm that augments diffusion models with low-dimensional latent variables that capture high-level factors of variation in the data. InfoDiffusion relies on a learning objective regularized with the mutual information between observed and hidden variables, which improves latent space quality and prevents the latents from being ignored by expressive diffusion-based decoders. Empirically, we find that InfoDiffusion learns disentangled and human-interpretable latent representations that are competitive with state-of-the-art generative and contrastive methods, while retaining the high sample quality of diffusion models. Our method enables manipulating the attributes of generated images and has the potential to assist tasks that require exploring a learned latent space to generate quality samples, e.g., generative design.

****

A Robust Test for the Stationarity Assumption in Sequential Decision Making
Jitao Wang, Chengchun Shi, Zhenke Wu
Reinforcement learning (RL) is a powerful technique that allows an autonomous ag ent to learn an optimal policy to maximize the expected return. The optimality o f various RL algorithms relies on the stationarity assumption, which requires ti me-invariant state transition and reward functions. However, deviations from sta tionarity over extended periods often occur in real-world applications like robo tics control, health care and digital marketing, resulting in suboptimal policie s learned under stationary assumptions. In this paper, we propose a model-based doubly robust procedure for testing the stationarity assumption and detecting ch ange points in offline RL settings with certain degree of homogeneity. Our propo sed testing procedure is robust to model misspecifications and can effectively c ontrol type-I error while achieving high statistical power, especially in high-d imensional settings. Extensive comparative simulations and a real-world interven tional mobile health example illustrate the advantages of our method in detectin g change points and optimizing long-term rewards in high-dimensional, non-statio nary environments.
****
GEAR: A GPU-Centric Experience Replay System for Large Reinforcement Learning Mo dels
Hanjing Wang, Man-Kit Sit, Congjie He, Ying Wen, Weinan Zhang, Jun Wang, Yaodong Yang, Luo Mai
This paper introduces a distributed, GPU-centric experience replay system, GEAR, designed to perform scalable reinforcement learning (RL) with large sequence mo dels (such as transformers). With such models, existing systems such as Reverb f ace considerable bottlenecks in memory, computation, and communication. GEAR, ho wever, optimizes memory efficiency by enabling the memory resources on GPU serve rs (including host memory and device memory) to manage trajectory data. Furtherm ore, it facilitates decentralized GPU devices to expedite various trajectory sel ection strategies, circumventing computational bottlenecks. GEAR is equipped wit h GPU kernels capable of collecting trajectories using zero-copy access to host memory, along with remote-directed-memory access over InfiniBand, improving comm unication efficiency. Cluster experiments have shown that GEAR can achieve perfo rmance levels up to 6× greater than Reverb when training state-of-the-art large RL models. GEAR is open-sourced at https:// github.com/bigrl-team/gear.
****
Effective and Efficient Structural Inference with Reservoir Computing
Aoran Wang, Tsz Pan Tong, Jun Pang
In this paper, we present an effective and efficient structural inference approa ch by integrating a Reservoir Computing (RC) network into a Variational Auto-enc oder-based (VAE-based) structural inference framework. With the help of Bi-level Optimization, the backbone VAE-based method follows the Information Bottleneck principle and infers a general adjacency matrix in its latent space; the RC net substitutes the partial role of the decoder and encourages the whole approach to perform further steps of gradient descent based on limited available data. The experimental results on various datasets including biological networks, simulate d fMRI data, and physical simulations show the effectiveness and efficiency of o ur proposed method for structural inference, either with much fewer trajectories or with much shorter trajectories compared with previous works.
****
Optimal Goal-Reaching Reinforcement Learning via Quasimetric Learning
Tongzhou Wang, Antonio Torralba, Phillip Isola, Amy Zhang
In goal-reaching reinforcement learning (RL), the optimal value function has a p articular geometry, called quasimetrics structure. This paper introduces Quasime tric Reinforcement Learning (QRL), a new RL method that utilizes quasimetric mod els to learn optimal value functions. Distinct from prior approaches, the QRL ob jective is specifically designed for quasimetrics, and provides strong theoretic al recovery guarantees. Empirically, we conduct thorough analyses on a discretiz ed MountainCar environment, identifying properties of QRL and its advantages ove r alternatives. On offline and online goal-reaching benchmarks, QRL also demonst

rates improved sample efficiency and performance, across both state-based and im
age-based observations.
****

Model-Free Robust Average-Reward Reinforcement Learning
Yue Wang, Alvaro Velasquez, George K. Atia, Ashley Prater-Bennette, Shaofeng Zou
Robust Markov decision processes (MDPs) address the challenge of model uncertain
ty by optimizing the worst-case performance over an uncertainty set of MDPs. In
this paper, we focus on the robust average-reward MDPs under the model-free sett
ing. We first theoretically characterize the structure of solutions to the robus
t average-reward Bellman equation, which is essential for our later convergence
analysis. We then design two model-free algorithms, robust relative value iterat
ion (RVI) TD and robust RVI Q-learning, and theoretically prove their convergenc
e to the optimal solution. We provide several widely used uncertainty sets as ex
amples, including those defined by the contamination model, total variation, Chi
-squared divergence, Kullback-Leibler (KL) divergence, and Wasserstein distance.
****

Live in the Moment: Learning Dynamics Model Adapted to Evolving Policy
Xiyao Wang, Wichayaporn Wongkamjan, Ruonan Jia, Furong Huang
Model-based reinforcement learning (RL) often achieves higher sample efficiency
in practice than model-free RL by learning a dynamics model to generate samples
for policy learning. Previous works learn a dynamics model that fits under the e
mpirical state-action visitation distribution for all historical policies, i.e.,
the sample replay buffer. However, in this paper, we observe that fitting the d
ynamics model under the distribution for all historical policies does not necess
arily benefit model prediction for the current policy since the policy in use is
constantly evolving over time. The evolving policy during training will cause s
tate-action visitation distribution shifts. We theoretically analyze how this di
stribution shift over historical policies affects the model learning and model r
ollouts. We then propose a novel dynamics model learning method, named Policy-ad
apted Dynamics Model Learning (PDML). PDML dynamically adjusts the historical po
licy mixture distribution to ensure the learned model can continually adapt to t
he state-action visitation distribution of the evolving policy. Experiments on a
range of continuous control environments in MuJoCo show that PDML achieves sign
ificant improvement in sample efficiency and higher asymptotic performance combi
ned with the state-of-the-art model-based RL methods.
****

Learning to Bid in Repeated First-Price Auctions with Budgets
Qian Wang, Zongjun Yang, Xiaotie Deng, Yuqing Kong
Budget management strategies in repeated auctions have received growing attentio
n in online advertising markets. However, previous work on budget management in
online bidding mainly focused on second-price auctions. The rapid shift from sec
ond-price auctions to first-price auctions for online ads in recent years has mo
tivated the challenging question of how to bid in repeated first-price auctions
while controlling budgets. In this work, we study the problem of learning in rep
eated first-price auctions with budgets. We design a dual-based algorithm that c
an achieve a near-optimal $\widetilde{O}(\sqrt{T})$ regret with full information
feedback where the maximum competing bid is always revealed after each auction.
We further consider the setting with one-sided information feedback where only
the winning bid is revealed after each auction. We show that our modified algori
thm can still achieve an $\widetilde{O}(\sqrt{T})$ regret with mild assumptions
on the bidder’s value distribution. Finally, we complement the theoretical resul
ts with numerical experiments to confirm the effectiveness of our budget managem
ent policy.
****

Network Effects in Performative Prediction Games
Xiaolu Wang, Chung-Yiu Yau, Hoi To Wai
This paper studies the multi-agent performative prediction (Multi-PP) games over
multiplex networks. We consider a distributed learning setting where agents par
tially cooperate on an agent network, while during learning, the data samples dr
awn depend on the prediction models of the agent itself and neighboring agents o

n a population network. The dynamics of Multi-PP games is hence affected by the interplay between both networks. This paper concentrates on this Multi-PP game with the following contributions. Firstly, we analyze sufficient conditions for the existence of the performative stable equilibrium (PSE) and Nash equilibrium (NE) of the Multi-PP games. Secondly, we analyze the changes to the equilibrium induced by perturbed data distributions, and derive the closed-form solutions where the network topologies are explicit. Our results connect the existence of PSE/NE with strengths of agents' cooperation, and the changes of equilibrium solutions across agents with their node centrality, etc. Lastly, we show that a stochastic gradient descent (SGD) based distributed learning procedure finds the PSE under the said sufficient condition. Numerical illustrations on the network effects in Multi-PP games corroborate our findings.

****

Robustly Learning a Single Neuron via Sharpness

Puqian Wang, Nikos Zarifis, Ilias Diakonikolas, Jelena Diakonikolas

We study the problem of learning a single neuron with respect to the L_2^2-loss in the presence of adversarial label noise. We give an efficient algorithm that, for a broad family of activations including ReLUs, approximates the optimal L_2^2-error within a constant factor. Notably, our algorithm succeeds under much milder distributional assumptions compared to prior work. The key ingredient enabling our results is a novel connection to local error bounds from optimization theory.

****

DualHSIC: HSIC-Bottleneck and Alignment for Continual Learning

Zifeng Wang, Zheng Zhan, Yifan Gong, Yucai Shao, Stratis Ioannidis, Yanzhi Wang, Jennifer Dy

Rehearsal-based approaches are a mainstay of continual learning (CL). They mitigate the catastrophic forgetting problem by maintaining a small fixed-size buffer with a subset of data from past tasks. While most rehearsal-based approaches exploit the knowledge from buffered past data, little attention is paid to inter-task relationships and to critical task-specific and task-invariant knowledge. By appropriately leveraging inter-task relationships, we propose a novel CL method, named DualHSIC, to boost the performance of existing rehearsal-based methods in a simple yet effective way. DualHSIC consists of two complementary components that stem from the so-called Hilbert Schmidt independence criterion (HSIC): HSIC-Bottleneck for Rehearsal (HBR) lessens the inter-task interference and HSIC Alignment (HA) promotes task-invariant knowledge sharing. Extensive experiments show that DualHSIC can be seamlessly plugged into existing rehearsal-based methods for consistent performance improvements, outperforming recent state-of-the-art regularization-enhanced rehearsal methods.

****

Enforcing Hard Constraints with Soft Barriers: Safe Reinforcement Learning in Unknown Stochastic Environments

Yixuan Wang, Simon Sinong Zhan, Ruochen Jiao, Zhilu Wang, Wanxin Jin, Zhuoran Yang, Zhaoran Wang, Chao Huang, Qi Zhu

It is quite challenging to ensure the safety of reinforcement learning (RL) agents in an unknown and stochastic environment under hard constraints that require the system state not to reach certain specified unsafe regions. Many popular safe RL methods such as those based on the Constrained Markov Decision Process (CMDP) paradigm formulate safety violations in a cost function and try to constrain the expectation of cumulative cost under a threshold. However, it is often difficult to effectively capture and enforce hard reachability-based safety constraints indirectly with such constraints on safety violation cost. In this work, we leverage the notion of barrier function to explicitly encode the hard safety chance constraints, and given that the environment is unknown, relax them to our design of generative-model-based soft barrier functions. Based on such soft barriers, we propose a novel safe RL approach with bi-level optimization that can jointly learn the unknown environment and optimize the control policy, while effectively avoiding the unsafe region with safety probability optimization. Experiments on a set of examples demonstrate that our approach can effectively enforce hard

safety chance constraints and significantly outperform CMDP-based baseline methods in system safe rates measured via simulations.
****
LinSATNet: The Positive Linear Satisfiability Neural Networks
Runzhong Wang, Yunhao Zhang, Ziao Guo, Tianyi Chen, Xiaokang Yang, Junchi Yan
Encoding constraints into neural networks is attractive. This paper studies how to introduce the popular positive linear satisfiability to neural networks. We propose the first differentiable satisfiability layer based on an extension of the classic Sinkhorn algorithm for jointly encoding multiple sets of marginal distributions. We further theoretically characterize the convergence property of the Sinkhorn algorithm for multiple marginals, and the underlying formulation is also derived. In contrast to the sequential decision e.g. reinforcement learning-based solvers, we showcase our technique in solving constrained (specifically satisfiability) problems by one-shot neural networks, including i) a neural routing solver learned without supervision of optimal solutions; ii) a partial graph matching network handling graphs with unmatchable outliers on both sides; iii) a predictive network for financial portfolios with continuous constraints. To our knowledge, there exists no one-shot neural solver for these scenarios when they are formulated as satisfiability problems. Source code is available at https://github.com/Thinklab-SJTU/LinSATNet.
****
Offline Meta Reinforcement Learning with In-Distribution Online Adaptation
Jianhao Wang, Jin Zhang, Haozhe Jiang, Junyu Zhang, Liwei Wang, Chongjie Zhang
Recent offline meta-reinforcement learning (meta-RL) methods typically utilize task-dependent behavior policies (e.g., training RL agents on each individual task) to collect a multi-task dataset. However, these methods always require extra information for fast adaptation, such as offline context for testing tasks. To address this problem, we first formally characterize a unique challenge in offline meta-RL: transition-reward distribution shift between offline datasets and online adaptation. Our theory finds that out-of-distribution adaptation episodes may lead to unreliable policy evaluation and that online adaptation with in-distribution episodes can ensure adaptation performance guarantee. Based on these theoretical insights, we propose a novel adaptation framework, called In-Distribution online Adaptation with uncertainty Quantification (IDAQ), which generates in-distribution context using a given uncertainty quantification and performs effective task belief inference to address new tasks. We find a return-based uncertainty quantification for IDAQ that performs effectively. Experiments show that IDAQ achieves state-of-the-art performance on the Meta-World ML1 benchmark compared to baselines with/without offline adaptation.
****
Reachability-Aware Laplacian Representation in Reinforcement Learning
Kaixin Wang, Kuangqi Zhou, Jiashi Feng, Bryan Hooi, Xinchao Wang
In Reinforcement Learning (RL), Laplacian Representation (LapRep) is a task-agnostic state representation that encodes the geometry of the environment. A desirable property of LapRep stated in prior works is that the Euclidean distance in the LapRep space roughly reflects the reachability between states, which motivates the usage of this distance for reward shaping. However, we find that LapRep does not necessarily have this property in general: two states having a small distance under LapRep can actually be far away in the environment. Such a mismatch would impede the learning process in reward shaping. To fix this issue, we introduce a Reachability-Aware Laplacian Representation (RA-LapRep), by properly scaling each dimension of LapRep. Despite the simplicity, we demonstrate that RA-LapRep can better capture the inter-state reachability as compared to LapRep, through both theoretical explanations and experimental results. Additionally, we show that this improvement yields a significant boost in reward shaping performance and benefits bottleneck state discovery.
****
PPG Reloaded: An Empirical Study on What Matters in Phasic Policy Gradient
Kaixin Wang, Daquan Zhou, Jiashi Feng, Shie Mannor
In model-free reinforcement learning, recent methods based on a phasic policy gr

adient (PPG) framework have shown impressive improvements in sample efficiency a
nd zero-shot generalization on the challenging Procgen benchmark. In PPG, two de
sign choices are believed to be the key contributing factors to its superior per
formance over PPO: the high level of value sample reuse and the low frequency of
 feature distillation. However, through an extensive empirical study, we unveil
that policy regularization and data diversity are what actually matters. In part
icular, we can achieve the same level of performance with low value sample reuse
 and frequent feature distillation, as long as the policy regularization strengt
h and data diversity are preserved. In addition, we can maintain the high perfor
mance of PPG while reducing the computational cost to a similar level as PPO. Ou
r comprehensive study covers all 16 Procgen games in both sample efficiency and
generalization setups. We hope it can advance the understanding of PPG and provi
de insights for future works.
****

On Heterogeneous Treatment Effects in Heterogeneous Causal Graphs
Richard A Watson, Hengrui Cai, Xinming An, Samuel Mclean, Rui Song
Heterogeneity and comorbidity are two interwoven challenges associated with vari
ous healthcare problems that greatly hampered research on developing effective t
reatment and understanding of the underlying neurobiological mechanism. Very few
 studies have been conducted to investigate heterogeneous causal effects (HCEs)
in graphical contexts due to the lack of statistical methods. To characterize th
is heterogeneity, we first conceptualize heterogeneous causal graphs (HCGs) by g
eneralizing the causal graphical model with confounder-based interactions and mu
ltiple mediators. Such confounders with an interaction with the treatment are kn
own as moderators. This allows us to flexibly produce HCGs given different moder
ators and explicitly characterize HCEs from the treatment or potential mediators
 on the outcome. We establish the theoretical forms of HCEs and derive their pro
perties at the individual level in both linear and nonlinear models. An interact
ive structural learning is developed to estimate the complex HCGs and HCEs with
confidence intervals provided. Our method is empirically justified by extensive
simulations and its practical usefulness is illustrated by exploring causality a
mong psychiatric disorders for trauma survivors. Code implementing the proposed
algorithm is open-source and publicly available at: https://github.com/richard-w
atson/ISL.
****

Nonparametric Extensions of Randomized Response for Private Confidence Sets
Ian Waudby-Smith, Steven Wu, Aaditya Ramdas
This work derives methods for performing nonparametric, nonasymptotic statistica
l inference for population means under the constraint of local differential priv
acy (LDP). Given bounded observations $(X_1, …, X_n)$ with mean $\mu^\star$ that
 are privatized into $(Z_1, …, Z_n)$, we present confidence intervals (CI) and t
ime-uniform confidence sequences (CS) for $\mu^\star$ when only given access to
the privatized data. To achieve this, we introduce a nonparametric and sequentia
lly interactive generalization of Warner’s famous “randomized response” mechanis
m, satisfying LDP for arbitrary bounded random variables, and then provide CIs a
nd CSs for their means given access to the resulting privatized observations. Fo
r example, our results yield private analogues of Hoeffding’s inequality in both
 fixed-time and time-uniform regimes. We extend these Hoeffding-type CSs to capt
ure time-varying (non-stationary) means, and conclude by illustrating how these
methods can be used to conduct private online A/B tests.
****

Global optimality for Euclidean CCCP under Riemannian convexity
Melanie Weber, Suvrit Sra
We study geodesically convex (g-convex) problems that can be written as a differ
ence of Euclidean convex functions. This structure arises in key applications su
ch as matrix scaling, M- estimators of scatter matrices, and Brascamp-Lieb inequ
alities. In particular, we exploit this structure to make use of the Convex-Conc
ave Procedure (CCCP), which helps us bypass potentially expensive Riemannian ope
rations and leads to very competitive solvers. Importantly, unlike existing theo
ry for CCCP that ensures convergence to stationary points, we exploit the overal

l g-convexity structure and provide iteration complexity results for global optimality. We illustrate our results by specializing them to a few concrete optimization problems that have been previously studied in the machine learning literature. We hope our work spurs the study of mixed Euclidean-Riemannian optimization algorithms.

****

A Universal Unbiased Method for Classification from Aggregate Observations

Zixi Wei, Lei Feng, Bo Han, Tongliang Liu, Gang Niu, Xiaofeng Zhu, Heng Tao Shen

In conventional supervised classification, true labels are required for individual instances. However, it could be prohibitive to collect the true labels for individual instances, due to privacy concerns or unaffordable annotation costs. This motivates the study on classification from aggregate observations (CFAO), where the supervision is provided to groups of instances, instead of individual instances. CFAO is a generalized learning framework that contains various learning problems, such as multiple-instance learning and learning from label proportions. The goal of this paper is to present a novel universal method of CFAO, which holds an unbiased estimator of the classification risk for arbitrary losses—previous research failed to achieve this goal. Practically, our method works by weighing the importance of each instance and each label in the group, which provides purified supervision for the classifier to learn. Theoretically, our proposed method not only guarantees the risk consistency due to the unbiased risk estimator but also can be compatible with arbitrary losses. Extensive experiments on various problems of CFAO demonstrate the superiority of our proposed method.

****

NTK-approximating MLP Fusion for Efficient Language Model Fine-tuning

Tianxin Wei, Zeming Guo, Yifan Chen, Jingrui He

Fine-tuning a pre-trained language model (PLM) emerges as the predominant strategy in many natural language processing applications. However, even fine-tuning the PLMs and doing inference are expensive, especially on edge devices with low computing power. Some general approaches (e.g. quantization and distillation) have been widely studied to reduce the compute/memory of PLM fine-tuning, while very few one-shot compression techniques are explored. In this paper, we investigate the neural tangent kernel (NTK)–which reveals the gradient descent dynamics of neural networks–of the multilayer perceptrons (MLP) modules in a PLM and propose to coin a lightweight PLM through NTK-approximating MLP fusion. To achieve this, we reconsider the MLP as a bundle of sub-MLPs, and cluster them into a given number of centroids, which can then be restored as a compressed MLP and surprisingly shown to well approximate the NTK of the original PLM. Extensive experiments of PLM fine-tuning on both natural language understanding (NLU) and generation (NLG) tasks are provided to verify the effectiveness of the proposed method MLP fusion. Our code is available at https://github.com/weitianxin/MLP_Fusion.

****

Boosting Graph Contrastive Learning via Graph Contrastive Saliency

Chunyu Wei, Yu Wang, Bing Bai, Kai Ni, David Brady, Lu Fang

Graph augmentation plays a crucial role in achieving good generalization for contrastive graph self-supervised learning. However, mainstream Graph Contrastive Learning (GCL) often favors random graph augmentations, by relying on random node dropout or edge perturbation on graphs. Random augmentations may inevitably lead to semantic information corruption during the training, and force the network to mistakenly focus on semantically irrelevant environmental background structures. To address these limitations and to improve generalization, we propose a novel self-supervised learning framework for GCL, which can adaptively screen the semantic-related substructure in graphs by capitalizing on the proposed gradient-based Graph Contrastive Saliency (GCS). The goal is to identify the most semantically discriminative structures of a graph via contrastive learning, such that we can generate semantically meaningful augmentations by leveraging on saliency. Empirical evidence on 16 benchmark datasets demonstrates the exclusive merits of the GCS-based framework. We also provide rigorous theoretical justification for GCS’s robustness properties. Code is available at https://github.com/GCS2023/GCS .

****
Set-membership Belief State-based Reinforcement Learning for POMDPs
Wei Wei, Lijun Zhang, Lin Li, Huizhong Song, Jiye Liang
Reinforcement learning (RL) has made significant progress in areas such as Atari games and robotic control, where the agents have perfect sensing capabilities. However, in many real-world sequential decision-making tasks, the observation data could be noisy or incomplete due to the intrinsic low quality of the sensors or unexpected malfunctions; that is, the agent's perceptions are rarely perfect. The current POMDP RL methods, such as particle-based and Gaussian-based, can only provide a probability estimate of hidden states rather than certain belief regions, which may lead to inefficient and even wrong decision-making. This paper proposes a novel algorithm called Set-membership Belief state-based Reinforcement Learning (SBRL), which consists of two parts: a Set-membership Belief state learning Model (SBM) for learning bounded belief state sets and an RL controller for making decisions based on SBM. We prove that our belief estimation method can provide a series of belief state sets that always contain the true states under the unknown-but-bounded (UBB) noise. The effectiveness of the proposed method is verified on a collection of benchmark tasks, and the results show that our method outperforms the state-of-the-art methods.
****
Mitigating Memorization of Noisy Labels by Clipping the Model Prediction
Hongxin Wei, Huiping Zhuang, Renchunzi Xie, Lei Feng, Gang Niu, Bo An, Yixuan Li
In the presence of noisy labels, designing robust loss functions is critical for securing the generalization performance of deep neural networks. Cross Entropy (CE) loss has been shown to be not robust to noisy labels due to its unboundedness. To alleviate this issue, existing works typically design specialized robust losses with the symmetric condition, which usually lead to the underfitting issue. In this paper, our key idea is to induce a loss bound at the logit level, thus universally enhancing the noise robustness of existing losses. Specifically, we propose logit clipping (LogitClip), which clamps the norm of the logit vector to ensure that it is upper bounded by a constant. In this manner, CE loss equipped with our LogitClip method is effectively bounded, mitigating the overfitting to examples with noisy labels. Moreover, we present theoretical analyses to certify the noise-tolerant ability of LogitClip. Extensive experiments show that LogitClip not only significantly improves the noise robustness of CE loss, but also broadly enhances the generalization performance of popular robust losses.
****
Graphically Structured Diffusion Models
Christian Dietrich Weilbach, William Harvey, Frank Wood
We introduce a framework for automatically defining and learning deep generative models with problem-specific structure. We tackle problem domains that are more traditionally solved by algorithms such as sorting, constraint satisfaction for Sudoku, and matrix factorization. Concretely, we train diffusion models with an architecture tailored to the problem specification. This problem specification should contain a graphical model describing relationships between variables, and often benefits from explicit representation of subcomputations. Permutation invariances can also be exploited. Across a diverse set of experiments we improve the scaling relationship between problem dimension and our model's performance, in terms of both training time and final accuracy. Our code can be found at https://github.com/plai-group/gsdm.
****
Expectation-Complete Graph Representations with Homomorphisms
Pascal Welke, Maximilian Thiessen, Fabian Jogl, Thomas Gärtner
We investigate novel random graph embeddings that can be computed in expected polynomial time and that are able to distinguish all non-isomorphic graphs in expectation. Previous graph embeddings have limited expressiveness and either cannot distinguish all graphs or cannot be computed efficiently for every graph. To be able to approximate arbitrary functions on graphs, we are interested in efficient alternatives that become arbitrarily expressive with increasing resources. Our approach is based on Lovász' characterisation of graph isomorphism through an

infinite dimensional vector of homomorphism counts. Our empirical evaluation sho ws competitive results on several benchmark graph learning tasks.
****
A Conditional Normalizing Flow for Accelerated Multi-Coil MR Imaging
Jeffrey Wen, Rizwan Ahmad, Philip Schniter
Accelerated magnetic resonance (MR) imaging attempts to reduce acquisition time by collecting data below the Nyquist rate. As an ill-posed inverse problem, many plausible solutions exist, yet the majority of deep learning approaches generat e only a single solution. We instead focus on sampling from the posterior distri bution, which provides more comprehensive information for downstream inference t asks. To do this, we design a novel conditional normalizing flow (CNF) that infe rs the signal component in the measurement operator's nullspace, which is later combined with measured data to form complete images. Using fastMRI brain and kne e data, we demonstrate fast inference and accuracy that surpasses recent posteri or sampling techniques for MRI. Code is available at https://github.com/jwen307/ mri_cnf
****
Optimizing Mode Connectivity for Class Incremental Learning
Haitao Wen, Haoyang Cheng, Heqian Qiu, Lanxiao Wang, Lili Pan, Hongliang Li
Class incremental learning (CIL) is one of the most challenging scenarios in con tinual learning. Existing work mainly focuses on strategies like memory replay, regularization, or dynamic architecture but ignores a crucial aspect: mode conne ctivity. Recent studies have shown that different minima can be connected by a l ow-loss valley, and ensembling over the valley shows improved performance and ro bustness. Motivated by this, we try to investigate the connectivity in CIL and f ind that the high-loss ridge exists along the linear connection between two adja cent continual minima. To dodge the ridge, we propose parameter-saving OPtimizin g Connectivity (OPC) based on Fourier series and gradient projection for finding the low-loss path between minima. The optimized path provides infinite low-loss solutions. We further propose EOPC to ensemble points within a local bent cylin der to improve performance on learned tasks. Our scheme can serve as a plug-in u nit, extensive experiments on CIFAR-100, ImageNet-100, and ImageNet-1K show cons istent improvements when adapting EOPC to existing representative CIL methods. O ur code is available at https://github.com/HaitaoWen/EOPC.
****
Towards Learning Geometric Eigen-Lengths Crucial for Fitting Tasks
Yijia Weng, Kaichun Mo, Ruoxi Shi, Yanchao Yang, Leonidas Guibas
Some extremely low-dimensional yet crucial geometric eigen-lengths often determi ne the success of some geometric tasks. For example, the height of an object is important to measure to check if it can fit between the shelves of a cabinet, wh ile the width of a couch is crucial when trying to move it through a doorway. Hu mans have materialized such crucial geometric eigen-lengths in common sense sinc e they are very useful in serving as succinct yet effective, highly interpretabl e, and universal object representations. However, it remains obscure and underex plored if learning systems can be equipped with similar capabilities of automati cally discovering such key geometric quantities from doing tasks. In this work, we therefore for the first time formulate and propose a novel learning problem o n this question and set up a benchmark suite including tasks, data, and evaluati on metrics for studying the problem. We focus on a family of common fitting task s as the testbed for the proposed learning problem. We explore potential solutio ns and demonstrate the feasibility of learning eigen-lengths from simply observi ng successful and failed fitting trials. We also attempt geometric grounding for more accurate eigen-length measurement and study the reusability of the learned geometric eigen-lengths across multiple tasks. Our work marks the first explora tory step toward learning crucial geometric eigen-lengths and we hope it can ins pire future research in tackling this important yet underexplored problem.
****
Open-VCLIP: Transforming CLIP to an Open-vocabulary Video Model via Interpolated Weight Optimization
Zejia Weng, Xitong Yang, Ang Li, Zuxuan Wu, Yu-Gang Jiang

Contrastive Language-Image Pretraining (CLIP) has demonstrated impressive zero-shot learning abilities for image understanding, yet limited effort has been made to investigate CLIP for zero-shot video recognition. We introduce Open-VCLIP, a simple yet effective approach that transforms CLIP into a strong zero-shot video classifier that can recognize unseen actions and events at test time. Our framework extends CLIP with minimal modifications to model spatial-temporal relationships in videos, making it a specialized video classifier, while striving for generalization. We formally show that training an Open-VCLIP is equivalent to continual learning with zero historical data. To address this problem, we propose Interpolated Weight Optimization, which utilizes the benefit of weight interpolation in both training and test time. We evaluate our method on three popular and challenging action recognition datasets following various zero-shot evaluation protocols and we demonstrate our approach outperforms state-of-the-art methods by clear margins. In particular, we achieve 87.9%, 58.3%, 81.1% zero-shot accuracy on UCF, HMDB and Kinetics-600 respectively, outperforming state-of-the-art methods by 8.3%, 7.8% and 12.2%. Code is released at https://github.com/wengzejia1/Open-VCLIP.

****

Fully-Adaptive Composition in Differential Privacy

Justin Whitehouse, Aaditya Ramdas, Ryan Rogers, Steven Wu

Composition is a key feature of differential privacy. Well-known advanced composition theorems allow one to query a private database quadratically more times than a basic privacy composition would permit. However, these results require that the privacy parameters of all algorithms be fixed before interacting with the data. To address this, Rogers et al. introduced fully adaptive composition, wherein both algorithms and their privacy parameters can be selected adaptively. They defined two probabilistic objects to measure privacy in adaptive composition: privacy filters, which provide differential privacy guarantees for composed interactions, and privacy odometers, time-uniform bounds on privacy loss. There are substantial gaps between advanced composition and existing filters and odometers. First, existing filters place stronger assumptions on the algorithms being composed. Second, these odometers and filters suffer from large constants, making them impractical. We construct filters that match the rates of advanced composition, including constants, despite allowing for adaptively chosen privacy parameters. En route we also derive a privacy filter for approximate zCDP. We also construct several general families of odometers. These odometers match the tightness of advanced composition at an arbitrary, preselected point in time, or at all points in time simultaneously, up to a doubly-logarithmic factor. We obtain our results by leveraging advances in martingale concentration. In sum, we show that fully adaptive privacy is obtainable at almost no loss.

****

Scalable Set Encoding with Universal Mini-Batch Consistency and Unbiased Full Set Gradient Approximation

Jeffrey Willette, Seanie Lee, Bruno Andreis, Kenji Kawaguchi, Juho Lee, Sung Ju Hwang

Recent work on mini-batch consistency (MBC) for set functions has brought attention to the need for sequentially processing and aggregating chunks of a partitioned set while guaranteeing the same output for all partitions. However, existing constraints on MBC architectures lead to models with limited expressive power. Additionally, prior work has not addressed how to deal with large sets during training when the full set gradient is required. To address these issues, we propose a Universally MBC (UMBC) class of set functions which can be used in conjunction with arbitrary non-MBC components while still satisfying MBC, enabling a wider range of function classes to be used in MBC settings. Furthermore, we propose an efficient MBC training algorithm which gives an unbiased approximation of the full set gradient and has a constant memory overhead for any set size for both train- and test-time. We conduct extensive experiments including image completion, text classification, unsupervised clustering, and cancer detection on high-resolution images to verify the efficiency and efficacy of our scalable set encoding framework. Our code is available at github.com/jeffwillette/umbc

****
Flexible Phase Dynamics for Bio-Plausible Contrastive Learning
Ezekiel Williams, Colin Bredenberg, Guillaume Lajoie
Many learning algorithms used as normative models in neuroscience or as candidate approaches for learning on neuromorphic chips learn by contrasting one set of network states with another. These Contrastive Learning (CL) algorithms are traditionally implemented with rigid, temporally non-local, and periodic learning dynamics, that could limit the range of physical systems capable of harnessing CL. In this study, we build on recent work exploring how CL might be implemented by biological or neurmorphic systems and show that this form of learning can be made temporally local, and can still function even if many of the dynamical requirements of standard training procedures are relaxed. Thanks to a set of general theorems corroborated by numerical experiments across several CL models, our results provide theoretical foundations for the study and development of CL methods for biological and neuromorphic neural networks.
****
Approximate Stein Classes for Truncated Density Estimation
Daniel James Williams, Song Liu
Estimating truncated density models is difficult, as these models have intractable normalising constants and hard to satisfy boundary conditions. Score matching can be adapted to solve the truncated density estimation problem, but requires a continuous weighting function which takes zero at the boundary and is positive elsewhere. Evaluation of such a weighting function (and its gradient) often requires a closed-form expression of the truncation boundary and finding a solution to a complicated optimisation problem. In this paper, we propose approximate Stein classes, which in turn leads to a relaxed Stein identity for truncated density estimation. We develop a novel discrepancy measure, truncated kernelised Stein discrepancy (TKSD), which does not require fixing a weighting function in advance, and can be evaluated using only samples on the boundary. We estimate a truncated density model by minimising the Lagrangian dual of TKSD. Finally, experiments show the accuracy of our method to be an improvement over previous works even without the explicit functional form of the boundary.
****
Theoretical Behavior of XAI Methods in the Presence of Suppressor Variables
Rick Wilming, Leo Kieslich, Benedict Clark, Stefan Haufe
In recent years, the community of ’explainable artificial intelligence’ (XAI) has created a vast body of methods to bridge a perceived gap between model ’complexity’ and ’interpretability’. However, a concrete problem to be solved by XAI methods has not yet been formally stated. As a result, XAI methods are lacking theoretical and empirical evidence for the ’correctness’ of their explanations, limiting their potential use for quality-control and transparency purposes. At the same time, Haufe et al. (2014) showed, using simple toy examples, that even standard interpretations of linear models can be highly misleading. Specifically, high importance may be attributed to so-called suppressor variables lacking any statistical relation to the prediction target. This behavior has been confirmed empirically for a large array of XAI methods in Wilming et al. (2022). Here, we go one step further by deriving analytical expressions for the behavior of a variety of popular XAI methods on a simple two-dimensional binary classification problem involving Gaussian class-conditional distributions. We show that the majority of the studied approaches will attribute non-zero importance to a non-class-related suppressor feature in the presence of correlated noise. This poses important limitations on the interpretations and conclusions that the outputs of these XAI methods can afford.
****
Marginalization is not Marginal: No Bad VAE Local Minima when Learning Optimal Sparse Representations
David Wipf
Although the variational autoencoder (VAE) represents a widely-used deep generative model, the underlying energy function when applied to continuous data remains poorly understood. In fact, most prior theoretical analysis has assumed a simp

lified affine decoder such that the model collapses to probabilistic PCA, a restricted regime whereby existing classical algorithms can also be trivially applied to guarantee globally optimal solutions. To push our understanding into more complex, practically-relevant settings, this paper instead adopts a deceptively sophisticated single-layer decoder that nonetheless allows the VAE to address the fundamental challenge of learning optimally sparse representations of continuous data originating from popular multiple-response regression models. In doing so, we can then examine VAE properties within the non-trivial context of solving difficult, NP-hard inverse problems. More specifically, we prove rigorous conditions which guarantee that any minimum of the VAE energy (local or global) will produce the optimally sparse latent representation, meaning zero reconstruction error using a minimal number of active latent dimensions. This is ultimately possible because VAE marginalization over the latent posterior selectively smooths away bad local minima as has been conjectured but not actually proven in prior work. We then discuss how equivalent-capacity deterministic autoencoders, even with appropriate sparsity-promoting regularization of the latent space, maintain bad local minima that do not correspond with such parsimonious representations. Overall, these results serve to elucidate key properties of the VAE loss surface relative to finding low-dimensional structure in data.

****

Uncertainty Estimation for Molecules: Desiderata and Methods

Tom Wollschläger, Nicholas Gao, Bertrand Charpentier, Mohamed Amine Ketata, Stephan Günnemann

Graph Neural Networks (GNNs) are promising surrogates for quantum mechanical calculations as they establish unprecedented low errors on collections of molecular dynamics (MD) trajectories. Thanks to their fast inference times they promise to accelerate computational chemistry applications. Unfortunately, despite low in-distribution (ID) errors, such GNNs might be horribly wrong for out-of-distribution (OOD) samples. Uncertainty estimation (UE) may aid in such situations by communicating the model’s certainty about its prediction. Here, we take a closer look at the problem and identify six key desiderata for UE in molecular force fields, three ’physics-informed’ and three ’application-focused’ ones. To overview the field, we survey existing methods from the field of UE and analyze how they fit to the set desiderata. By our analysis, we conclude that none of the previous works satisfies all criteria. To fill this gap, we propose Localized Neural Kernel (LNK) a Gaussian Process (GP)-based extension to existing GNNs satisfying the desiderata. In our extensive experimental evaluation, we test four different UE with three different backbones across two datasets. In out-of-equilibrium detection, we find LNK yielding up to 2.5 and 2.1 times lower errors in terms of AUC-ROC score than dropout or evidential regression-based methods while maintaining high predictive performance.

****

The Blessing of Heterogeneity in Federated Q-Learning: Linear Speedup and Beyond

Jiin Woo, Gauri Joshi, Yuejie Chi

In this paper, we consider federated Q-learning, which aims to learn an optimal Q-function by periodically aggregating local Q-estimates trained on local data alone. Focusing on infinite-horizon tabular Markov decision processes, we provide sample complexity guarantees for both the synchronous and asynchronous variants of federated Q-learning. In both cases, our bounds exhibit a linear speedup with respect to the number of agents and sharper dependencies on other salient problem parameters. Moreover, existing approaches to federated Q-learning adopt an equally-weighted averaging of local Q-estimates, which can be highly sub-optimal in the asynchronous setting since the local trajectories can be highly heterogeneous due to different local behavior policies. Existing sample complexity scales inverse proportionally to the minimum entry of the stationary state-action occupancy distributions over all agents, requiring that every agent covers the entire state-action space. Instead, we propose a novel importance averaging algorithm, giving larger weights to more frequently visited state-action pairs. The improved sample complexity scales inverse proportionally to the minimum entry of the average stationary state-action occupancy distribution of all agents, thus only

requiring the agents collectively cover the entire state-action space, unveiling the blessing of heterogeneity.
****
Learning Deep Time-index Models for Time Series Forecasting
Gerald Woo, Chenghao Liu, Doyen Sahoo, Akshat Kumar, Steven Hoi
Deep learning has been actively applied to time series forecasting, leading to a deluge of new methods, belonging to the class of historical-value models. Yet, despite the attractive properties of time-index models, such as being able to model the continuous nature of underlying time series dynamics, little attention has been given to them. Indeed, while naive deep time-index models are far more expressive than the manually predefined function representations of classical time-index models, they are inadequate for forecasting, being unable to generalize to unseen time steps due to the lack of inductive bias. In this paper, we propose DeepTime, a meta-optimization framework to learn deep time-index models which overcome these limitations, yielding an efficient and accurate forecasting model. Extensive experiments on real world datasets in the long sequence time-series forecasting setting demonstrate that our approach achieves competitive results with state-of-the-art methods, and is highly efficient. Code is available at https://github.com/salesforce/DeepTime.
****
Sharper Bounds for ℓ_p Sensitivity Sampling
David Woodruff, Taisuke Yasuda
In large scale machine learning, random sampling is a popular way to approximate datasets by a small representative subset of examples. In particular, sensitivity sampling is an intensely studied technique which provides provable guarantees on the quality of approximation, while reducing the number of examples to the product of the VC dimension d and the total sensitivity $\mathfrak{S}$ in remarkably general settings. However, guarantees going beyond this general bound of $\mathfrak{S} d$ are known in perhaps only one setting, for ℓ_2 subspace embeddings, despite intense study of sensitivity sampling in prior work. In this work, we show the first bounds for sensitivity sampling for ℓ_p subspace embeddings for $p\neq 2$ that improve over the general $\mathfrak{S} d$ bound, achieving a bound of roughly $\mathfrak{S}^{2/p}$ for $1\leq p<2$ and $\mathfrak{S}^{2-2/p}$ for $2root leverage score sampling algorithm achieves a bound of roughly d for $1\leq p<2$, and that a combination of leverage score and sensitivity sampling achieves an improved bound of roughly $d^{2/p}\mathfrak{S}^{2-4/p}$ for $2
Cite this Paper

BibTeX

@InProceedings{pmlr-v202-woodruff23a,
  title = ■ {Sharper Bounds for ℓ_p Sensitivity Sampling},
  author =   {Woodruff, David and Yasuda, Taisuke},
  booktitle = ■ {Proceedings of the 40th International Conference on Machine Learning},
  pages = ■ {37238--37272},
  year = ■ {2023},
  editor = ■ {Krause, Andreas and Brunskill, Emma and Cho, Kyunghyun and Engelhardt, Barbara and Sabato, Sivan and Scarlett, Jonathan},
  volume = ■ {202},
  series = ■ {Proceedings of Machine Learning Research},
  month = ■ {23--29 Jul},
  publisher =   {PMLR},
  pdf = ■ {https://proceedings.mlr.press/v202/woodruff23a/woodruff23a.pdf},

```
  url = ■ {https://proceedings.mlr.press/v202/woodruff23a.html},
  abstract = ■ {In large scale machine learning, random sampling is a popular wa
y to approximate datasets by a small representative subset of examples. In parti
cular, sensitivity sampling is an intensely studied technique which provides pro
vable guarantees on the quality of approximation, while reducing the number of e
xamples to the product of the VC dimension $d$ and the total sensitivity $\mathf
rak{S}$ in remarkably general settings. However, guarantees going beyond this ge
neral bound of $\mathfrak{S} d$ are known in perhaps only one setting, for $\ell
_2$ subspace embeddings, despite intense study of sensitivity sampling in prior
work. In this work, we show the first bounds for sensitivity sampling for $\ell_
p$ subspace embeddings for $p\neq 2$ that improve over the general $\mathfrak{S}
 d$ bound, achieving a bound of roughly $\mathfrak{S}^{2/p}$ for $1\leq p<2$ and
 $\mathfrak{S}^{2-2/p}$ for $2root leverage score sampling algorithm achieves a
bound of roughly $d$ for $1\leq p<2$, and that a combination of leverage score a
nd sensitivity sampling achieves an improved bound of roughly $d^{2/p}\mathfrak{
S}^{2-4/p}$ for $2

Copy to Clipboard
Download



    Endnote



%0 Conference Paper
%T Sharper Bounds for $\ell_p$ Sensitivity Sampling
%A David Woodruff
%A Taisuke Yasuda
%B Proceedings of the 40th International Conference on Machine Learning
%C Proceedings of Machine Learning Research
%D 2023
%E Andreas Krause
%E Emma Brunskill
%E Kyunghyun Cho
%E Barbara Engelhardt
%E Sivan Sabato
%E Jonathan Scarlett■
%F pmlr-v202-woodruff23a
%I PMLR
%P 37238--37272
%U https://proceedings.mlr.press/v202/woodruff23a.html
%V 202
%X In large scale machine learning, random sampling is a popular way to approxim
ate datasets by a small representative subset of examples. In particular, sensit
ivity sampling is an intensely studied technique which provides provable guarant
ees on the quality of approximation, while reducing the number of examples to th
e product of the VC dimension $d$ and the total sensitivity $\mathfrak{S}$ in re
markably general settings. However, guarantees going beyond this general bound o
f $\mathfrak{S} d$ are known in perhaps only one setting, for $\ell_2$ subspace
embeddings, despite intense study of sensitivity sampling in prior work. In this
 work, we show the first bounds for sensitivity sampling for $\ell_p$ subspace e
mbeddings for $p\neq 2$ that improve over the general $\mathfrak{S} d$ bound, ac
hieving a bound of roughly $\mathfrak{S}^{2/p}$ for $1\leq p<2$ and $\mathfrak{S
}^{2-2/p}$ for $2root leverage score sampling algorithm achieves a bound of roug
hly $d$ for $1\leq p<2$, and that a combination of leverage score and sensitivit
y sampling achieves an improved bound of roughly $d^{2/p}\mathfrak{S}^{2-4/p}$ f
or $2
```

Copy to Clipboard
Download

APA

Woodruff, D. & Yasuda, T.. (2023). Sharper Bounds for ℓ_p Sensitivity Sampling. Proceedings of the 40th International Conference on Machine Learning, in Proceedings of Machine Learning Research 202:37238-37272 Available from https://proceedings.mlr.press/v202/woodruff23a.html.

Copy to Clipboard
Download

Related Material

Download PDF
OpenReview
****
Two Losses Are Better Than One: Faster Optimization Using a Cheaper Proxy
Blake Woodworth, Konstantin Mishchenko, Francis Bach
We present an algorithm for minimizing an objective with hard-to-compute gradients by using a related, easier-to-access function as a proxy. Our algorithm is based on approximate proximal-point iterations on the proxy combined with relatively few stochastic gradients from the objective. When the difference between the objective and the proxy is δ-smooth, our algorithm guarantees convergence at a rate matching stochastic gradient descent on a δ-smooth objective, which can lead to substantially better sample efficiency. Our algorithm has many potential applications in machine learning, and provides a principled means of leveraging synthetic data, physics simulators, mixed public and private data, and more.
****
SEGA: Structural Entropy Guided Anchor View for Graph Contrastive Learning
Junran Wu, Xueyuan Chen, Bowen Shi, Shangzhe Li, Ke Xu
In contrastive learning, the choice of "view" controls the information that the representation captures and influences the performance of the model. However, leading graph contrastive learning methods generally produce views via random corruption or learning, which could lead to the loss of essential information and alteration of semantic information. An anchor view that maintains the essential information of input graphs for contrastive learning has been hardly investigated. In this paper, based on the theory of graph information bottleneck, we deduce the definition of this anchor view; put differently, the anchor view with essential information of input graph is supposed to have the minimal structural uncertainty. Furthermore, guided by structural entropy, we implement the anchor view, termed SEGA, for graph contrastive learning. We extensively validate the proposed anchor view on various benchmarks regarding graph classification under unsupervised, semi-supervised, and transfer learning and achieve significant performance boosts compared to the state-of-the-art methods.
****

Causal Proxy Models for Concept-based Model Explanations
Zhengxuan Wu, Karel D’Oosterlinck, Atticus Geiger, Amir Zur, Christopher Potts
Explainability methods for NLP systems encounter a version of the fundamental pr oblem of causal inference: for a given ground-truth input text, we never truly o bserve the counterfactual texts necessary for isolating the causal effects of mo del representations on outputs. In response, many explainability methods make no use of counterfactual texts, assuming they will be unavailable. In this paper, we show that robust causal explainability methods can be created using approxima te counterfactuals, which can be written by humans to approximate a specific cou nterfactual or simply sampled using metadata-guided heuristics. The core of our proposal is the Causal Proxy Model (CPM). A CPM explains a black-box model $\mat hcal{N}$ because it is trained to have the same actual input/output behavior as $\mathcal{N}$ while creating neural representations that can be intervened upon to simulate the counterfactual input/output behavior of $\mathcal{N}$. Furthermo re, we show that the best CPM for $\mathcal{N}$ performs comparably to $\mathcal {N}$ in making factual predictions, which means that the CPM can simply replace $\mathcal{N}$, leading to more explainable deployed models.
****
Effective Neural Topic Modeling with Embedding Clustering Regularization
Xiaobao Wu, Xinshuai Dong, Thong Thanh Nguyen, Anh Tuan Luu
Topic models have been prevalent for decades with various applications. However, existing topic models commonly suffer from the notorious topic collapsing: disc overed topics semantically collapse towards each other, leading to highly repeti tive topics, insufficient topic discovery, and damaged model interpretability. I n this paper, we propose a new neural topic model, Embedding Clustering Regulari zation Topic Model (ECRTM). Besides the existing reconstruction error, we propos e a novel Embedding Clustering Regularization (ECR), which forces each topic emb edding to be the center of a separately aggregated word embedding cluster in the semantic space. This enables each produced topic to contain distinct word seman tics, which alleviates topic collapsing. Regularized by ECR, our ECRTM generates diverse and coherent topics together with high-quality topic distributions of d ocuments. Extensive experiments on benchmark datasets demonstrate that ECRTM eff ectively addresses the topic collapsing issue and consistently surpasses state-o f-the-art baselines in terms of topic quality, topic distributions of documents, and downstream classification tasks.
****
Adaptive Compositional Continual Meta-Learning
Bin Wu, Jinyuan Fang, Xiangxiang Zeng, Shangsong Liang, Qiang Zhang
This paper focuses on continual meta-learning, where few-shot tasks are heteroge neous and sequentially available. Recent works use a mixture model for meta-know ledge to deal with the heterogeneity. However, these methods suffer from paramet er inefficiency caused by two reasons: (1) the underlying assumption of mutual e xclusiveness among mixture components hinders sharing meta-knowledge across hete rogeneous tasks. (2) they only allow increasing mixture components and cannot ad aptively filter out redundant components. In this paper, we propose an Adaptive Compositional Continual Meta-Learning (ACML) algorithm, which employs a composit ional premise to associate a task with a subset of mixture components, allowing meta-knowledge sharing among heterogeneous tasks. Moreover, to adaptively adjust the number of mixture components, we propose a component sparsification method based on evidential theory to filter out redundant components. Experimental resu lts show ACML outperforms strong baselines, showing the effectiveness of our com positional meta-knowledge, and confirming that ACML can adaptively learn meta-kn owledge.
****
Anchor Sampling for Federated Learning with Partial Client Participation
Feijie Wu, Song Guo, Zhihao Qu, Shiqi He, Ziming Liu, Jing Gao
Compared with full client participation, partial client participation is a more practical scenario in federated learning, but it may amplify some challenges in federated learning, such as data heterogeneity. The lack of inactive clients’ up dates in partial client participation makes it more likely for the model aggrega

tion to deviate from the aggregation based on full client participation. Trainin g with large batches on individual clients is proposed to address data heterogen eity in general, but their effectiveness under partial client participation is n ot clear. Motivated by these challenges, we propose to develop a novel federated learning framework, referred to as FedAMD, for partial client participation. Th e core idea is anchor sampling, which separates partial participants into anchor and miner groups. Each client in the anchor group aims at the local bullseye wi th the gradient computation using a large batch. Guided by the bullseyes, client s in the miner group steer multiple near-optimal local updates using small batch es and update the global model. By integrating the results of the two groups, Fe dAMD is able to accelerate the training process and improve the model performanc e. Measured by ϵ-approximation and compared to the state-of-the-art met hods, FedAMD achieves the convergence by up to $O(1/\epsilon)$ fewer communicati on rounds under non-convex objectives. Empirical studies on real-world datasets validate the effectiveness of FedAMD and demonstrate the superiority of the prop osed algorithm: Not only does it considerably save computation and communication costs, but also the test accuracy significantly improves.

****

Solving High-Dimensional PDEs with Latent Spectral Models

Haixu Wu, Tengge Hu, Huakun Luo, Jianmin Wang, Mingsheng Long

Deep models have achieved impressive progress in solving partial differential eq uations (PDEs). A burgeoning paradigm is learning neural operators to approximat e the input-output mappings of PDEs. While previous deep models have explored th e multiscale architectures and various operator designs, they are limited to lea rning the operators as a whole in the coordinate space. In real physical science problems, PDEs are complex coupled equations with numerical solvers relying on discretization into high-dimensional coordinate space, which cannot be precisely approximated by a single operator nor efficiently learned due to the curse of d imensionality. We present Latent Spectral Models (LSM) toward an efficient and p recise solver for high-dimensional PDEs. Going beyond the coordinate space, LSM enables an attention-based hierarchical projection network to reduce the high-di mensional data into a compact latent space in linear time. Inspired by classical spectral methods in numerical analysis, we design a neural spectral block to so lve PDEs in the latent space that approximates complex input-output mappings via learning multiple basis operators, enjoying nice theoretical guarantees for con vergence and approximation. Experimentally, LSM achieves consistent state-of-the -art and yields a relative gain of 11.5% averaged on seven benchmarks covering b oth solid and fluid physics. Code is available at https://github.com/thuml/Laten t-Spectral-Models.

****

A Law of Robustness beyond Isoperimetry

Yihan Wu, Heng Huang, Hongyang Zhang

We study the robust interpolation problem of arbitrary data distributions suppor ted on a bounded space and propose a two-fold law of robustness. Robust interpol ation refers to the problem of interpolating n noisy training data points in $ R^d$ by a Lipschitz function. Although this problem has been well understood whe n the samples are drawn from an isoperimetry distribution, much remains unknown concerning its performance under generic or even the worst-case distributions. W e prove a Lipschitzness lower bound $\Omega(\sqrt{n/p})$ of the interpolating ne ural network with p parameters on arbitrary data distributions. With this resu lt, we validate the law of robustness conjecture in prior work by Bubeck, Li and Nagaraj on two-layer neural networks with polynomial weights. We then extend ou r result to arbitrary interpolating approximators and prove a Lipschitzness lowe r bound $\Omega(n^{1/d})$ for robust interpolation. Our results demonstrate a tw o-fold law of robustness: a) we show the potential benefit of overparametrizatio n for smooth data interpolation when $n=poly(d)$, and b) we disprove the potenti al existence of an $O(1)$-Lipschitz robust interpolating function when $n=\exp(\ omega(d))$.

****

Uncovering Adversarial Risks of Test-Time Adaptation

Tong Wu, Feiran Jia, Xiangyu Qi, Jiachen T. Wang, Vikash Sehwag, Saeed Mahloujifar, Prateek Mittal

Recently, test-time adaptation (TTA) has been proposed as a promising solution for addressing distribution shifts. It allows a base model to adapt to an unforeseen distribution during inference by leveraging the information from the batch of (unlabeled) test data. However, we uncover a novel security vulnerability of TTA based on the insight that predictions on benign samples can be impacted by malicious samples in the same batch. To exploit this vulnerability, we propose Distribution Invading Attack (DIA), which injects a small fraction of malicious data into the test batch. DIA causes models using TTA to misclassify benign and unperturbed test data, providing an entirely new capability for adversaries that is infeasible in canonical machine learning pipelines. Through comprehensive evaluations, we demonstrate the high effectiveness of our attack on multiple benchmarks across six TTA methods. In response, we investigate two countermeasures to robustify the existing insecure TTA implementations, following the principle of security by design. Together, we hope our findings can make the community aware of the utility-security tradeoffs in deploying TTA and provide valuable insights for developing robust TTA approaches.

****

Stable Estimation of Heterogeneous Treatment Effects

Anpeng Wu, Kun Kuang, Ruoxuan Xiong, Bo Li, Fei Wu

Estimating heterogeneous treatment effects (HTE) is crucial for identifying the variation of treatment effects across individuals or subgroups. Most existing methods estimate HTE by removing the confounding bias from imbalanced treatment assignments. However, these methods may produce unreliable estimates of treatment effects and potentially allocate suboptimal treatment arms for underrepresented populations. To improve the estimation accuracy of HTE for underrepresented populations, we propose a novel Stable CounterFactual Regression (StableCFR) to smooth the population distribution and upsample the underrepresented subpopulations, while balancing confounders between treatment and control groups. Specifically, StableCFR upsamples the underrepresented data using uniform sampling, where each disjoint subpopulation is weighted proportional to the Lebesgue measure of its support. Moreover, StableCFR balances covariates by using an epsilon-greedy matching approach. Empirical results on both synthetic and real-world datasets demonstrate the superior performance of our StableCFR on estimating HTE for underrepresented populations.

****

Rethinking Explaining Graph Neural Networks via Non-parametric Subgraph Matching

Fang Wu, Siyuan Li, Xurui Jin, Yinghui Jiang, Dragomir Radev, Zhangming Niu, Stan Z. Li

The success of graph neural networks (GNNs) provokes the question about explainability: “Which fraction of the input graph is the most determinant of the prediction?” Particularly, parametric explainers prevail in existing approaches because of their more robust capability to decipher the black-box (i.e., target GNNs). In this paper, based on the observation that graphs typically share some common motif patterns, we propose a novel non-parametric subgraph matching framework, dubbed MatchExplainer, to explore explanatory subgraphs. It couples the target graph with other counterpart instances and identifies the most crucial joint substructure by minimizing the node corresponding-based distance. Moreover, we note that present graph sampling or node-dropping methods usually suffer from the false positive sampling problem. To alleviate this issue, we design a new augmentation paradigm named MatchDrop. It takes advantage of MatchExplainer to fix the most informative portion of the graph and merely operates graph augmentations on the rest less informative part. Extensive experiments on synthetic and real-world datasets show the effectiveness of our MatchExplainer by outperforming all state-of-the-art parametric baselines with significant margins. Results also demonstrate that MatchDrop is a general scheme to be equipped with GNNs for enhanced performance. The code is available at https://github.com/smiles724/MatchExplainer.

****

Understanding Int4 Quantization for Language Models: Latency Speedup, Composabil

ity, and Failure Cases
Xiaoxia Wu, Cheng Li, Reza Yazdani Aminabadi, Zhewei Yao, Yuxiong He
Improving the deployment efficiency of transformer-based language models has bee n challenging given their high computation and memory cost. While INT8 quantizat ion has recently been shown to be effective in reducing both the memory cost and latency while preserving model accuracy, it remains unclear whether we can leve rage INT4 (which doubles peak hardware throughput) to achieve further latency im provement. In this study, we explore the feasibility of employing INT4 weight an d activation (W4A4) quantization for language models. Our findings indicate that W4A4 quantization introduces no to negligible accuracy degradation for encoder- only and encoder-decoder models, but causes a significant accuracy drop for deco der-only models. To materialize the performance gain using W4A4, we develop a hi ghly-optimized end-to-end W4A4 encoder inference pipeline supporting different q uantization strategies. Our INT4 pipeline is $8.5\times$ faster for latency-orie nted scenarios and up to $3\times$ for throughput-oriented scenarios compared to the inference of FP16, and improves the SOTA BERT INT8 performance from FasterT ransformer by up to $1.7\times$. We provide insights into the failure cases when applying W4A4 to decoder-only models, and further explore the compatibility of INT4 quantization with other compression methods, like pruning and layer reducti on.
****

Towards Understanding Generalization of Macro-AUC in Multi-label Learning
Guoqiang Wu, Chongxuan Li, Yilong Yin
Macro-AUC is the arithmetic mean of the class-wise AUCs in multi-label learning and is commonly used in practice. However, its theoretical understanding is far lacking. Toward solving it, we characterize the generalization properties of var ious learning algorithms based on the corresponding surrogate losses w.r.t. Macr o-AUC. We theoretically identify a critical factor of the dataset affecting the generalization bounds: the label-wise class imbalance. Our results on the imbala nce-aware error bounds show that the widely-used univariate loss-based algorithm is more sensitive to the label-wise class imbalance than the proposed pairwise and reweighted loss-based ones, which probably implies its worse performance. Mo reover, empirical results on various datasets corroborate our theory findings. T o establish it, technically, we propose a new (and more general) McDiarmid-type concentration inequality, which may be of independent interest.
****

Quantifying the Knowledge in GNNs for Reliable Distillation into MLPs
Lirong Wu, Haitao Lin, Yufei Huang, Stan Z. Li
To bridge the gaps between topology-aware Graph Neural Networks (GNNs) and infer ence-efficient Multi-Layer Perceptron (MLPs), GLNN proposes to distill knowledge from a well-trained teacher GNN into a student MLP. Despite their great progres s, comparatively little work has been done to explore the reliability of differe nt knowledge points (nodes) in GNNs, especially their roles played during distil lation. In this paper, we first quantify the knowledge reliability in GNN by mea suring the invariance of their information entropy to noise perturbations, from which we observe that different knowledge points (1) show different distillation speeds (temporally); (2) are differentially distributed in the graph (spatially). To achieve reliable distillation, we propose an effective approach, namely Kn owledge-inspired Reliable Distillation (KRD), that models the probability of eac h node being an informative and reliable knowledge point, based on which we samp le a set of additional reliable knowledge points as supervision for training stu dent MLPs. Extensive experiments show that KRD improves over the vanilla MLPs by 12.62% and outperforms its corresponding teacher GNNs by 2.16% averaged over 7 datasets and 3 GNN architectures. Codes are publicly available at: https://githu b.com/LirongWu/RKD.
****

Delay-agnostic Asynchronous Coordinate Update Algorithm
Xuyang Wu, Changxin Liu, Sindri Magnússon, Mikael Johansson
We propose a delay-agnostic asynchronous coordinate update algorithm (DEGAS) for computing operator fixed points, with applications to asynchronous optimization

. DEGAS includes novel asynchronous variants of ADMM and block-coordinate descen t as special cases. We prove that DEGAS converges with both bounded and unbounde d delays under delay-free parameter conditions. We also validate by theory and e xperiments that DEGAS adapts well to the actual delays. The effectiveness of DEG AS is demonstrated by numerical experiments on classification problems.
****

Masked Trajectory Models for Prediction, Representation, and Control
Philipp Wu, Arjun Majumdar, Kevin Stone, Yixin Lin, Igor Mordatch, Pieter Abbeel , Aravind Rajeswaran
We introduce Masked Trajectory Models (MTM) as a generic abstraction for sequent ial decision making. MTM takes a trajectory, such as a state-action sequence, an d aims to reconstruct the trajectory conditioned on random subsets of the same t rajectory. By training with a highly randomized masking pattern, MTM learns vers atile networks that can take on different roles or capabilities, by simply choos ing appropriate masks at inference time. For example, the same MTM network can b e used as a forward dynamics model, inverse dynamics model, or even an offline R L agent. Through extensive experiments in several continuous control tasks, we s how that the same MTM network – i.e. same weights – can match or outperform spec ialized networks trained for the aforementioned capabilities. Additionally, we f ind that state representations learned by MTM can significantly accelerate the l earning speed of traditional RL algorithms. Finally, in offline RL benchmarks, w e find that MTM is competitive with specialized offline RL algorithms, despite M TM being a generic self-supervised learning method without any explicit RL compo nents. Code is available at https://github.com/facebookresearch/mtm.
****

Disentangled Multi-Fidelity Deep Bayesian Active Learning
Dongxia Wu, Ruijia Niu, Matteo Chinazzi, Yian Ma, Rose Yu
To balance quality and cost, various domain areas of science and engineering run simulations at multiple levels of sophistication. Multi-fidelity active learnin g aims to learn a direct mapping from input parameters to simulation outputs at the highest fidelity by actively acquiring data from multiple fidelity levels. H owever, existing approaches based on Gaussian processes are hardly scalable to h igh-dimensional data. Deep learning-based methods often impose a hierarchical st ructure in hidden representations, which only supports passing information from low-fidelity to high-fidelity. These approaches can lead to the undesirable prop agation of errors from low-fidelity representations to high-fidelity ones. We pr opose a novel framework called Disentangled Multi-fidelity Deep Bayesian Active Learning (D-MFDAL), which learns the surrogate models conditioned on the distrib ution of functions at multiple fidelities. On benchmark tasks of learning deep s urrogates of partial differential equations including heat equation, Poisson’s e quation and fluid simulations, our approach significantly outperforms state-of-t he-art in prediction accuracy and sample efficiency.
****

Tight Data Access Bounds for Private Top-k Selection
Hao Wu, Olga Ohrimenko, Anthony Wirth
We study the top-k selection problem under the differential privacy model: m items are rated according to votes of a set of clients. We consider a setting i n which algorithms can retrieve data via a sequence of accesses, each either a r andom access or a sorted access; the goal is to minimize the total number of dat a accesses. Our algorithm requires only $O(\sqrt{mk})$ expected accesses: to our knowledge, this is the first sublinear data-access upper bound for this problem . Our analysis also shows that the well-known exponential mechanism requires onl y $O(\sqrt{m})$ expected accesses. Accompanying this, we develop the first lower bounds for the problem, in three settings: only random accesses; only sorted ac cesses; a sequence of accesses of either kind. We show that, to avoid $\Omega(m) $ access cost, supporting both kinds of access is necessary, and that in this ca se our algorithm’s access cost is optimal.
****

The Implicit Regularization of Dynamical Stability in Stochastic Gradient Descen t

Lei Wu, Weijie J Su
In this paper, we study the implicit regularization of stochastic gradient descent (SGD) through the lens of dynamical stability (Wu et al., 2018). We start by revising existing stability analyses of SGD, showing how the Frobenius norm and trace of Hessian relate to different notions of stability. Notably, if a global minimum is linearly stable for SGD, then the trace of Hessian must be less than or equal to $2/\eta$, where η denotes the learning rate. By contrast, for gradient descent (GD), the stability imposes a similar constraint but only on the largest eigenvalue of Hessian. We then turn to analyze the generalization properties of these stable minima, focusing specifically on two-layer ReLU networks and diagonal linear networks. Notably, we establish the equivalence between these metrics of sharpness and certain parameter norms for the two models, which allows us to show that the stable minima of SGD provably generalize well. By contrast, the stability-induced regularization of GD is provably too weak to ensure satisfactory generalization. This discrepancy provides an explanation of why SGD often generalizes better than GD. Note that the learning rate (LR) plays a pivotal role in the strength of stability-induced regularization. As the LR increases, the regularization effect becomes more pronounced, elucidating why SGD with a larger LR consistently demonstrates superior generalization capabilities. Additionally, numerical experiments are provided to support our theoretical findings.
****

Distributional Offline Policy Evaluation with Predictive Error Guarantees
Runzhe Wu, Masatoshi Uehara, Wen Sun
We study the problem of estimating the distribution of the return of a policy using an offline dataset that is not generated from the policy, i.e., distributional offline policy evaluation (OPE). We propose an algorithm called Fitted Likelihood Estimation (FLE), which conducts a sequence of Maximum Likelihood Estimation (MLE) and has the flexibility of integrating any state-of-the-art probabilistic generative models as long as it can be trained via MLE. FLE can be used for both finite-horizon and infinite-horizon discounted settings where rewards can be multi-dimensional vectors. Our theoretical results show that for both finite-horizon and infinite-horizon discounted settings, FLE can learn distributions that are close to the ground truth under total variation distance and Wasserstein distance, respectively. Our theoretical results hold under the conditions that the offline data covers the test policy’s traces and that the supervised learning MLE procedures succeed. Experimentally, we demonstrate the performance of FLE with two generative models, Gaussian mixture models and diffusion models. For the multi-dimensional reward setting, FLE with diffusion models is capable of estimating the complicated distribution of the return of a test policy.
****

π-Tuning: Transferring Multimodal Foundation Models with Optimal Multi-task Interpolation
Chengyue Wu, Teng Wang, Yixiao Ge, Zeyu Lu, Ruisong Zhou, Ying Shan, Ping Luo
Foundation models have achieved great advances in multi-task learning with a unified interface of unimodal and multimodal tasks. However, the potential of such multi-task learners has not been exploited during transfer learning. In this work, we present a universal parameter-efficient transfer learning method, termed Predict-Interpolate Tuning (π-Tuning), for vision, language, and vision-language tasks. It aggregates the parameters of lightweight task-specific experts learned from similar tasks to aid the target downstream task. The task similarities are predicted in a unified modality-independent space, yielding a scalable graph to demonstrate task relationships. π-Tuning has several appealing benefits. First, it flexibly explores both intra- and inter-modal transferability between similar tasks to improve the accuracy and robustness of transfer learning, especially in data-scarce scenarios. Second, it offers a systematical solution for transfer learning with multi-task prediction-and-then-interpolation, compatible with diverse types of parameter-efficient experts, such as prompt and adapter. Third, an extensive study of task-level mutual benefits on 14 unimodal and 6 multimodal datasets shows that π-Tuning surpasses fine-tuning and other parameter-efficient transfer learning methods both in full-shot and low-shot regimes. Th

e task graph also enables an in-depth interpretable analysis of task transferability across modalities. The code will be available at https://github.com/TencentARC/pi-Tuning.

****

Learning Functional Distributions with Private Labels

Changlong Wu, Yifan Wang, Ananth Grama, Wojciech Szpankowski

We study the problem of learning functional distributions in the presence of noise. A functional is a map from the space of features to distributions over a set of labels, and is often assumed to belong to a known class of hypotheses $\mathcal{F}$. Features are generated by a general random process and labels are sampled independently from feature-dependent distributions. In privacy sensitive applications, labels are passed through a noisy kernel. We consider online learning, where at each time step, a predictor attempts to predict the actual (label) distribution given only the features and noisy labels in prior steps. The performance of the predictor is measured by the expected KL-risk that compares the predicted distributions to the underlying truth. We show that the minimax expected KL-risk is of order $\tilde{\Theta}(\sqrt{T\log|\mathcal{F}|})$ for finite hypothesis class $\mathcal{F}$ and any non-trivial noise level. We then extend this result to general infinite classes via the concept of stochastic sequential covering and provide matching lower and upper bounds for a wide range of natural classes.

****

QuantumDARTS: Differentiable Quantum Architecture Search for Variational Quantum Algorithms

Wenjie Wu, Ge Yan, Xudong Lu, Kaisen Pan, Junchi Yan

With the arrival of the Noisy Intermediate-Scale Quantum (NISQ) era and the fast development of machine learning, variational quantum algorithms (VQA) including Variational Quantum Eigensolver (VQE) and quantum neural network (QNN) have received increasing attention with wide potential applications in foreseeable near future. We study the problem of quantum architecture search (QAS) for VQA to automatically design parameterized quantum circuits (PQC). We devise a differentiable searching algorithm based on Gumbel-Softmax in contrast to peer methods that often require numerous circuit sampling and evaluation. Two versions of our algorithm are provided, namely macro search and micro search, where macro search directly searches for the whole circuit like other literature while the innovative micro search is able to infer the sub-circuit structure from a small-scale and then transfer that to a large-scale problem. We conduct intensive experiments on unweighted Max-Cut, ground state energy estimation, and image classification. The superior performance shows the efficiency and capability of macro search, which requires little prior knowledge. Moreover, the experiments on micro search show the potential of our algorithm for large-scale QAS problems.

****

Discover and Cure: Concept-aware Mitigation of Spurious Correlation

Shirley Wu, Mert Yuksekgonul, Linjun Zhang, James Zou

Deep neural networks often rely on spurious correlations to make predictions, which hinders generalization beyond training environments. For instance, models that associate cats with bed backgrounds can fail to predict the existence of cats in other environments without beds. Mitigating spurious correlations is crucial in building trustworthy models. However, the existing works lack transparency to offer insights into the mitigation process. In this work, we propose an interpretable framework, Discover and Cure (DISC), to tackle the issue. With human-interpretable concepts, DISC iteratively 1) discovers unstable concepts across different environments as spurious attributes, then 2) intervenes on the training data using the discovered concepts to reduce spurious correlation. Across systematic experiments, DISC provides superior generalization ability and interpretability than the existing approaches. Specifically, it outperforms the state-of-the-art methods on an object recognition task and a skin-lesion classification task by 7.5% and 9.6%, respectively. Additionally, we offer theoretical analysis and guarantees to understand the benefits of models trained by DISC. Code and data are available at https://github.com/Wuyxin/DISC.

****
On the Training Instability of Shuffling SGD with Batch Normalization
David Xing Wu, Chulhee Yun, Suvrit Sra
We uncover how SGD interacts with batch normalization and can exhibit undesirable training dynamics such as divergence. More precisely, we study how Single Shuffle (SS) and Random Reshuffle (RR)—two widely used variants of SGD—interact surprisingly differently in the presence of batch normalization: RR leads to much more stable evolution of training loss than SS. As a concrete example, for regression using a linear network with batch normalized inputs, we prove that SS and RR converge to distinct global optima that are “distorted” away from gradient descent. Thereafter, for classification we characterize conditions under which training divergence for SS and RR can, and cannot occur. We present explicit constructions to show how SS leads to distorted optima in regression and divergence for classification, whereas RR avoids both distortion and divergence. We validate our results empirically in realistic settings, and conclude that the separation between SS and RR used with batch normalization is relevant in practice.
****
dugMatting: Decomposed-Uncertainty-Guided Matting
Jiawei Wu, Changqing Zhang, Zuoyong Li, Huazhu Fu, Xi Peng, Joey Tianyi Zhou
Cutting out an object and estimating its opacity mask, known as image matting, is a key task in image and video editing. Due to the highly ill-posed issue, additional inputs, typically user-defined trimaps or scribbles, are usually needed to reduce the uncertainty. Although effective, it is either time consuming or only suitable for experienced users who know where to place the strokes. In this work, we propose a decomposed-uncertainty-guided matting (dugMatting) algorithm, which explores the explicitly decomposed uncertainties to efficiently and effectively improve the results. Basing on the characteristic of these uncertainties, the epistemic uncertainty is reduced in the process of guiding interaction (which introduces prior knowledge), while the aleatoric uncertainty is reduced in modeling data distribution (which introduces statistics for both data and possible noise). The proposed matting framework relieves the requirement for users to determine the interaction areas by using simple and efficient labeling. Extensively quantitative and qualitative results validate that the proposed method significantly improves the original matting algorithms in terms of both efficiency and efficacy.
****
Personalized Federated Learning under Mixture of Distributions
Yue Wu, Shuaicheng Zhang, Wenchao Yu, Yanchi Liu, Quanquan Gu, Dawei Zhou, Haifeng Chen, Wei Cheng
The recent trend towards Personalized Federated Learning (PFL) has garnered significant attention as it allows for the training of models that are tailored to each client while maintaining data privacy. However, current PFL techniques primarily focus on modeling the conditional distribution heterogeneity (i.e. concept shift), which can result in suboptimal performance when the distribution of input data across clients diverges (i.e. covariate shift). Additionally, these techniques often lack the ability to adapt to unseen data, further limiting their effectiveness in real-world scenarios. To address these limitations, we propose a novel approach, FedGMM, which utilizes Gaussian mixture models (GMM) to effectively fit the input data distributions across diverse clients. The model parameters are estimated by maximum likelihood estimation utilizing a federated Expectation-Maximization algorithm, which is solved in closed form and does not assume gradient similarity. Furthermore, FedGMM possesses an additional advantage of adapting to new clients with minimal overhead, and it also enables uncertainty quantification. Empirical evaluations on synthetic and benchmark datasets demonstrate the superior performance of our method in both PFL classification and novel sample detection.
****
Differentially Private Episodic Reinforcement Learning with Heavy-tailed Rewards
Yulian Wu, Xingyu Zhou, Sayak Ray Chowdhury, Di Wang
In this paper we study the problem of (finite horizon tabular) Markov decision p

rocesses (MDPs) with heavy-tailed rewards under the constraint of differential p
rivacy (DP). Compared with the previous studies for private reinforcement learni
ng that typically assume rewards are sampled from some bounded or sub-Gaussian d
istributions to ensure DP, we consider the setting where reward distributions ha
ve only finite $(1+v)$-th moments with some $v \in (0,1]$. By resorting to robus
t mean estimators for rewards, we first propose two frameworks for heavy-tailed
MDPs, i.e., one is for value iteration and another is for policy optimization. U
nder each framework, we consider both joint differential privacy (JDP) and local
 differential privacy (LDP) models. Based on our frameworks, we provide regret u
pper bounds for both JDP and LDP cases, and show that the moment of distribution
s and privacy budget have significant impact on regrets. Finally, we establish a
 lower bound of regret minimization for heavy-tailed MDPs in JDP model by reduci
ng it to the instance-independent lower bound of heavy-tailed multi-armed bandit
s in DP model. We also show the lower bound for the problem in LDP by adopting s
ome private minimax methods. Our results reveal that there are fundamental diffe
rences between the problem of private RL with sub-Gaussian and that with heavy-t
ailed rewards.
****

Finite-Sample Analysis of Learning High-Dimensional Single ReLU Neuron
Jingfeng Wu, Difan Zou, Zixiang Chen, Vladimir Braverman, Quanquan Gu, Sham M. K
akade
This paper considers the problem of learning single ReLU neuron with squared los
s (a.k.a., ReLU regression) in the overparameterized regime, where the input dim
ension can exceed the number of samples. We analyze a Perceptron-type algorithm
called GLM-tron [Kakade et al. 2011], and provide its dimension-free risk upper
bounds for high-dimensional ReLU regression in both well-specified and misspecif
ied settings. Our risk bounds recover several existing results as special cases.
 Moreover, in the well-specified setting, we also provide an instance-wise match
ing risk lower bound for GLM-tron. Our upper and lower risk bounds provide a sha
rp characterization of the high-dimensional ReLU regression problems that can be
 learned via GLM-tron. On the other hand, we provide some negative results for s
tochastic gradient descent (SGD) for ReLU regression with symmetric Bernoulli da
ta: if the model is well-specified, the excess risk of SGD is provably no better
 than that of GLM-tron ignoring constant factors, for each problem instance; and
 in the noiseless case, GLM-tron can achieve a small risk while SGD unavoidably
suffers from a constant risk in expectation. These results together suggest that
 GLM-tron might be more preferable than SGD for high-dimensional ReLU regression
.
****

Understanding Backdoor Attacks through the Adaptability Hypothesis
Xun Xian, Ganghua Wang, Jayanth Srinivasa, Ashish Kundu, Xuan Bi, Mingyi Hong, J
ie Ding
A poisoning backdoor attack is a rising security concern for deep learning. This
 type of attack can result in the backdoored model functioning normally most of
the time but exhibiting abnormal behavior when presented with inputs containing
the backdoor trigger, making it difficult to detect and prevent. In this work, w
e propose the adaptability hypothesis to understand when and why a backdoor atta
ck works for general learning models, including deep neural networks, based on t
he theoretical investigation of classical kernel-based learning models. The adap
tability hypothesis postulates that for an effective attack, the effect of incor
porating a new dataset on the predictions of the original data points will be sm
all, provided that the original data points are distant from the new dataset. Ex
periments on benchmark image datasets and state-of-the-art backdoor attacks for
deep neural networks are conducted to corroborate the hypothesis. Our finding pr
ovides insight into the factors that affect the attack’s effectiveness and has i
mplications for the design of future attacks and defenses.
****

Fair and Optimal Classification via Post-Processing
Ruicheng Xian, Lang Yin, Han Zhao
To mitigate the bias exhibited by machine learning models, fairness criteria can

be integrated into the training process to ensure fair treatment across all demographics, but it often comes at the expense of model performance. Understanding such tradeoffs, therefore, underlies the design of fair algorithms. To this end, this paper provides a complete characterization of the inherent tradeoff of demographic parity on classification problems, under the most general multi-group, multi-class, and noisy setting. Specifically, we show that the minimum error rate achievable by randomized and attribute-aware fair classifiers is given by the optimal value of a Wasserstein-barycenter problem. On the practical side, our findings lead to a simple post-processing algorithm that derives fair classifiers from score functions, which yields the optimal fair classifier when the score is Bayes optimal. We provide suboptimality analysis and sample complexity for our algorithm, and demonstrate its effectiveness on benchmark datasets.

****

UMD: Unsupervised Model Detection for X2X Backdoor Attacks

Zhen Xiang, Zidi Xiong, Bo Li

Backdoor (Trojan) attack is a common threat to deep neural networks, where samples from one or more source classes embedded with a backdoor trigger will be misclassified to adversarial target classes. Existing methods for detecting whether a classifier is backdoor attacked are mostly designed for attacks with a single adversarial target (e.g., all-to-one attack). To the best of our knowledge, without supervision, no existing methods can effectively address the more general X2X attack with an arbitrary number of source classes, each paired with an arbitrary target class. In this paper, we propose UMD, the first Unsupervised Model Detection method that effectively detects X2X backdoor attacks via a joint inference of the adversarial (source, target) class pairs. In particular, we first define a novel transferability statistic to measure and select a subset of putative backdoor class pairs based on a proposed clustering approach. Then, these selected class pairs are jointly assessed based on an aggregation of their reverse-engineered trigger size for detection inference, using a robust and unsupervised anomaly detector we proposed. We conduct comprehensive evaluations on CIFAR-10, GTSRB, and Imagenette dataset, and show that our unsupervised UMD outperforms SOTA detectors (even with supervision) by 17%, 4%, and 8%, respectively, in terms of the detection accuracy against diverse X2X attacks. We also show the strong detection performance of UMD against several strong adaptive attacks.

****

Random Shuffle Transformer for Image Restoration

Jie Xiao, Xueyang Fu, Man Zhou, Hongjian Liu, Zheng-Jun Zha

Non-local interactions play a vital role in boosting performance for image restoration. However, local window Transformer has been preferred due to its efficiency for processing high-resolution images. The superiority in efficiency comes at the cost of sacrificing the ability to model non-local interactions. In this paper, we present that local window Transformer can also function as modeling non-local interactions. The counterintuitive function is based on the permutation-equivariance of self-attention. The basic principle is quite simple: by randomly shuffling the input, local self-attention also has the potential to model non-local interactions without introducing extra parameters. Our random shuffle strategy enjoys elegant theoretical guarantees in extending the local scope. The resulting Transformer dubbed ShuffleFormer is capable of processing high-resolution images efficiently while modeling non-local interactions. Extensive experiments demonstrate the effectiveness of ShuffleFormer across a variety of image restoration tasks, including image denoising, deraining, and deblurring. Code is available at https://github.com/jiexiaou/ShuffleFormer.

****

Communication-Efficient Federated Hypergradient Computation via Aggregated Iterative Differentiation

Peiyao Xiao, Kaiyi Ji

Federated bilevel optimization has attracted increasing attention due to emerging machine learning and communication applications. The biggest challenge lies in computing the gradient of the upper-level objective function (i.e., hypergradient) in the federated setting due to the nonlinear and distributed construction o

f a series of global Hessian matrices. In this paper, we propose a novel communi cation-efficient federated hypergradient estimator via aggregated iterative diff erentiation (AggITD). AggITD is simple to implement and significantly reduces th e communication cost by conducting the federated hypergradient estimation and th e lower-level optimization simultaneously. We show that the proposed AggITD-base d algorithm achieves the same sample complexity as existing approximate implicit differentiation (AID)-based approaches with much fewer communication rounds in the presence of data heterogeneity. Our results also shed light on the great adv antage of ITD over AID in the federated/distributed hypergradient estimation. Th is differs from the comparison in the non-distributed bilevel optimization, wher e ITD is less efficient than AID. Our extensive experiments demonstrate the grea t effectiveness and communication efficiency of the proposed method.

****

SmoothQuant: Accurate and Efficient Post-Training Quantization for Large Languag e Models

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, Song Han

Large language models (LLMs) show excellent performance but are compute- and mem ory-intensive. Quantization can reduce memory and accelerate inference. However, existing methods cannot maintain accuracy and hardware efficiency at the same t ime. We propose SmoothQuant, a training-free, accuracy-preserving, and general-p urpose post-training quantization (PTQ) solution to enable 8-bit weight, 8-bit a ctivation (W8A8) quantization for LLMs. Based on the fact that weights are easy to quantize while activations are not, SmoothQuant smooths the activation outlie rs by offline migrating the quantization difficulty from activations to weights with a mathematically equivalent transformation. SmoothQuant enables an INT8 qua ntization of both weights and activations for all the matrix multiplications in LLMs, including OPT, BLOOM, GLM, MT-NLG, and LLaMA family. We demonstrate up to 1.56$\times$ speedup and 2$\times$ memory reduction for LLMs with negligible los s in accuracy. SmoothQuant enables serving 530B LLM within a single node. Our wo rk offers a turn-key solution that reduces hardware costs and democratizes LLMs.

****

On the Forward Invariance of Neural ODEs

Wei Xiao, Tsun-Hsuan Wang, Ramin Hasani, Mathias Lechner, Yutong Ban, Chuang Gan , Daniela Rus

We propose a new method to ensure neural ordinary differential equations (ODEs) satisfy output specifications by using invariance set propagation. Our approach uses a class of control barrier functions to transform output specifications int o constraints on the parameters and inputs of the learning system. This setup al lows us to achieve output specification guarantees simply by changing the constr ained parameters/inputs both during training and inference. Moreover, we demonst rate that our invariance set propagation through data-controlled neural ODEs not only maintains generalization performance but also creates an additional degree of robustness by enabling causal manipulation of the system’s parameters/inputs . We test our method on a series of representation learning tasks, including mod eling physical dynamics and convexity portraits, as well as safe collision avoid ance for autonomous vehicles.

****

COMCAT: Towards Efficient Compression and Customization of Attention-Based Visio n Models

Jinqi Xiao, Miao Yin, Yu Gong, Xiao Zang, Jian Ren, Bo Yuan

Attention-based vision models, such as Vision Transformer (ViT) and its variants , have shown promising performance in various computer vision tasks. However, th ese emerging architectures suffer from large model sizes and high computational costs, calling for efficient model compression solutions. To date, pruning ViTs has been well studied, while other compression strategies that have been widely applied in CNN compression, e.g., model factorization, is little explored in the context of ViT compression. This paper explores an efficient method for compres sing vision transformers to enrich the toolset for obtaining compact attention-b ased vision models. Based on the new insight on the multi-head attention layer, we develop a highly efficient ViT compression solution, which outperforms the st

ate-of-the-art pruning methods. For compressing DeiT-small and DeiT-base models on ImageNet, our proposed approach can achieve $0.45%$ and $0.76%$ higher top-1 accuracy even with fewer parameters. Our finding can also be applied to improve the customization efficiency of text-to-image diffusion models, with much faster training (up to $2.6\times$ speedup) and lower extra storage cost (up to $1927.5\times$ reduction) than the existing works.

****

Improving Bi-level Optimization Based Methods with Inspiration from Humans' Classroom Study Techniques

Pengtao Xie

In humans' classroom learning, many effective study techniques (e.g., the Feynman technique, peer questioning, etc.) have been developed to improve learning outcomes. We are interested in investigating whether these techniques can inspire the development of ML training strategies to improve bi-level optimization (BLO) based methods. Towards this goal, we develop a general framework, Skillearn, which consists of basic elements such as learners, interaction functions, learning stages, etc. These elements can be flexibly configured to create various training strategies, each emulating a study technique of humans. In case studies, we apply Skillearn to create new training strategies, by emulating the Feynman technique and peer questioning, which are two broadly adopted techniques in humans' classroom learning. These training strategies are used for improving two BLO based applications including neural architecture search and data weighting. Experiments on various datasets demonstrate the effectiveness of our methods.

****

Future-conditioned Unsupervised Pretraining for Decision Transformer

Zhihui Xie, Zichuan Lin, Deheng Ye, Qiang Fu, Yang Wei, Shuai Li

Recent research in offline reinforcement learning (RL) has demonstrated that return-conditioned supervised learning is a powerful paradigm for decision-making problems. While promising, return conditioning is limited to training data labeled with rewards and therefore faces challenges in learning from unsupervised data. In this work, we aim to utilize generalized future conditioning to enable efficient unsupervised pretraining from reward-free and sub-optimal offline data. We propose Pretrained Decision Transformer (PDT), a conceptually simple approach for unsupervised RL pretraining. PDT leverages future trajectory information as a privileged context to predict actions during training. The ability to make decisions based on both present and future factors enhances PDT's capability for generalization. Besides, this feature can be easily incorporated into a return-conditioned framework for online finetuning, by assigning return values to possible futures and sampling future embeddings based on their respective values. Empirically, PDT outperforms or performs on par with its supervised pretraining counterpart, especially when dealing with sub-optimal data. Further analysis reveals that PDT can extract diverse behaviors from offline data and controllably sample high-return behaviors by online finetuning. Code is available at here.

****

DeSRA: Detect and Delete the Artifacts of GAN-based Real-World Super-Resolution Models

Liangbin Xie, Xintao Wang, Xiangyu Chen, Gen Li, Ying Shan, Jiantao Zhou, Chao Dong

Image super-resolution (SR) with generative adversarial networks (GAN) has achieved great success in restoring realistic details. However, it is notorious that GAN-based SR models will inevitably produce unpleasant and undesirable artifacts, especially in practical scenarios. Previous works typically suppress artifacts with an extra loss penalty in the training phase. They only work for in-distribution artifact types generated during training. When applied in real-world scenarios, we observe that those improved methods still generate obviously annoying artifacts during inference. In this paper, we analyze the cause and characteristics of the GAN artifacts produced in unseen test data without ground-truths. We then develop a novel method, namely, DeSRA, to Detect and then "Delete" those SR Artifacts in practice. Specifically, we propose to measure a relative local variance distance from MSE-SR results and GAN-SR results, and locate the problematic

areas based on the above distance and semantic-aware thresholds. After detecting the artifact regions, we develop a finetune procedure to improve GAN-based SR models with a few samples, so that they can deal with similar types of artifacts in more unseen real data. Equipped with our DeSRA, we can successfully eliminate artifacts from inference and improve the ability of SR models to be applied in real-world scenarios. The code will be available at https://github.com/TencentARC/DeSRA.

****

Semiparametrically Efficient Off-Policy Evaluation in Linear Markov Decision Processes

Chuhan Xie, Wenhao Yang, Zhihua Zhang

We study semiparametrically efficient estimation in off-policy evaluation (OPE) where the underlying Markov decision process (MDP) is linear with a known feature map. We characterize the variance lower bound for regular estimators in the linear MDP setting and propose an efficient estimator whose variance achieves that lower bound. Consistency and asymptotic normality of our estimator are established under mild conditions, which merely requires the only infinite-dimensional nuisance parameter to be estimated at a $n^{-1/4}$ convergence rate. We also construct an asymptotically valid confidence interval for statistical inference and conduct simulation studies to validate our results. To our knowledge, this is the first work that concerns efficient estimation in the presence of a known structure of MDPs in the OPE literature.

****

A Critical View of Vision-Based Long-Term Dynamics Prediction Under Environment Misalignment

Hanchen Xie, Jiageng Zhu, Mahyar Khayatkhoei, Jiazhi Li, Mohamed E. Hussein, Wael Abdalmageed

Dynamics prediction, which is the problem of predicting future states of scene objects based on current and prior states, is drawing increasing attention as an instance of learning physics. To solve this problem, Region Proposal Convolutional Interaction Network (RPCIN), a vision-based model, was proposed and achieved state-of-the-art performance in long-term prediction. RPCIN only takes raw images and simple object descriptions, such as the bounding box and segmentation mask of each object, as input. However, despite its success, the model's capability can be compromised under conditions of environment misalignment. In this paper, we investigate two challenging conditions for environment misalignment: Cross-Domain and Cross-Context by proposing four datasets that are designed for these challenges: SimB-Border, SimB-Split, BlenB-Border, and BlenB-Split. The datasets cover two domains and two contexts. Using RPCIN as a probe, experiments conducted on the combinations of the proposed datasets reveal potential weaknesses of the vision-based long-term dynamics prediction model. Furthermore, we propose a promising direction to mitigate the Cross-Domain challenge and provide concrete evidence supporting such a direction, which provides dramatic alleviation of the challenge on the proposed datasets.

****

Controlling Type Confounding in Ad Hoc Teamwork with Instance-wise Teammate Feedback Rectification

Dong Xing, Pengjie Gu, Qian Zheng, Xinrun Wang, Shanqi Liu, Longtao Zheng, Bo An, Gang Pan

Ad hoc teamwork requires an agent to cooperate with unknown teammates without prior coordination. Many works propose to abstract teammate instances into high-level representation of types and then pre-train the best response for each type. However, most of them do not consider the distribution of teammate instances within a type. This could expose the agent to the hidden risk of type confounding. In the worst case, the best response for an abstract teammate type could be the worst response for all specific instances of that type. This work addresses the issue from the lens of causal inference. We first theoretically demonstrate that this phenomenon is due to the spurious correlation brought by uncontrolled teammate distribution. Then, we propose our solution, CTCAT, which disentangles such correlation through an instance-wise teammate feedback rectification. This oper

ation reweights the interaction of teammate instances within a shared type to re
duce the influence of type confounding. The effect of CTCAT is evaluated in mult
iple domains, including classic ad hoc teamwork tasks and real-world scenarios.
Results show that CTCAT is robust to the influence of type confounding, a practi
cal issue that directly hazards the robustness of our trained agents but was unn
oticed in previous works.
****

Universal Morphology Control via Contextual Modulation
Zheng Xiong, Jacob Beck, Shimon Whiteson
Learning a universal policy across different robot morphologies can significantl
y improve learning efficiency and generalization in continuous control. However,
 it poses a challenging multi-task reinforcement learning problem, as the optima
l policy may be quite different across robots and critically depend on the morph
ology. Existing methods utilize graph neural networks or transformers to handle
heterogeneous state and action spaces across different morphologies, but pay lit
tle attention to the dependency of a robot’s control policy on its morphology co
ntext. In this paper, we propose a hierarchical architecture to better model thi
s dependency via contextual modulation, which includes two key submodules: (1) I
nstead of enforcing hard parameter sharing across robots, we use hypernetworks t
o generate morphology-dependent control parameters; (2) We propose a fixed atten
tion mechanism that solely depends on the morphology to modulate the interaction
s between different limbs in a robot. Experimental results show that our method
not only improves learning performance on a diverse set of training robots, but
also generalizes better to unseen morphologies in a zero-shot fashion. The code
is publicly available at https://github.com/MasterXiong/ModuMorph.
****

Relevant Walk Search for Explaining Graph Neural Networks
Ping Xiong, Thomas Schnake, Michael Gastegger, Grégoire Montavon, Klaus Robert M
uller, Shinichi Nakajima
Graph Neural Networks (GNNs) have become important machine learning tools for gr
aph analysis, and its explainability is crucial for safety, fairness, and robust
ness. Layer-wise relevance propagation for GNNs (GNN-LRP) evaluates the relevanc
e of walks to reveal important information flows in the network, and provides hi
gher-order explanations, which have been shown to be superior to the lower-order
, i.e., node-/edge-level, explanations. However, identifying relevant walks by G
NN-LRP requires exponential computational complexity with respect to the network
 depth, which we will remedy in this paper. Specifically, we propose polynomial-
time algorithms for finding top-K relevant walks, which drastically reduces th
e computation and thus increases the applicability of GNN-LRP to large-scale pro
blems. Our proposed algorithms are based on the max-product algorithm—a common t
ool for finding the maximum likelihood configurations in probabilistic graphical
 models—and can find the most relevant walks exactly at the neuron level and app
roximately at the node level. Our experiments demonstrate the performance of our
 algorithms at scale and their utility across application domains, i.e., on epid
emiology, molecular, and natural language benchmarks. We provide our codes under
 github.com/xiong-ping/rel_walk_gnnlrp.
****

Why do Nearest Neighbor Language Models Work?
Frank F. Xu, Uri Alon, Graham Neubig
Language models (LMs) compute the probability of a text by sequentially computin
g a representation of an already-seen context and using this representation to p
redict the next word. Currently, most LMs calculate these representations throug
h a neural network consuming the immediate previous context. However recently, r
etrieval-augmented LMs have shown to improve over standard neural LMs, by access
ing information retrieved from a large datastore, in addition to their standard,
 parametric, next-word prediction. In this paper, we set out to understand why r
etrieval-augmented language models, and specifically why k-nearest neighbor lang
uage models (kNN-LMs) perform better than standard parametric LMs, even when the
 k-nearest neighbor component retrieves examples from the same training set that
 the LM was originally trained on. To this end, we perform analysis of various d

imensions over which kNN-LM diverges from standard LMs, and investigate these di
mensions one by one. Empirically, we identify three main reasons why kNN-LM perf
orms better than standard LMs: using a different input representation for predic
ting the next tokens, approximate kNN search, and the importance of softmax temp
erature for the kNN distribution. Further, we incorporate some insights into the
 standard parametric LM, improving performance without the need for an explicit
retrieval component. The code is available at https://github.com/frankxu2004/knn
lm-why.
****
MixFlows: principled variational inference via mixed flows
Zuheng Xu, Naitong Chen, Trevor Campbell
This work presents mixed variational flows (MixFlows), a new variational family
that consists of a mixture of repeated applications of a map to an initial refer
ence distribution. First, we provide efficient algorithms for i.i.d. sampling, d
ensity evaluation, and unbiased ELBO estimation. We then show that MixFlows have
 MCMC-like convergence guarantees when the flow map is ergodic and measure-prese
rving, and provide bounds on the accumulation of error for practical implementat
ions where the flow map is approximated. Finally, we develop an implementation o
f MixFlows based on uncorrected discretized Hamiltonian dynamics combined with d
eterministic momentum refreshment. Simulated and real data experiments show that
 MixFlows can provide more reliable posterior approximations than several black-
box normalizing flows, as well as samples of comparable quality to those obtaine
d from state-of-the-art MCMC methods.
****
Bit Allocation using Optimization
Tongda Xu, Han Gao, Chenjian Gao, Yuanyuan Wang, Dailan He, Jinyong Pi, Jixiang
Luo, Ziyu Zhu, Mao Ye, Hongwei Qin, Yan Wang, Jingjing Liu, Ya-Qin Zhang
In this paper, we consider the problem of bit allocation in Neural Video Compres
sion (NVC). First, we reveal a fundamental relationship between bit allocation i
n NVC and Semi-Amortized Variational Inference (SAVI). Specifically, we show tha
t SAVI with GoP (Group-of-Picture)-level likelihood is equivalent to pixel-level
 bit allocation with precise rate & quality dependency model. Based on this equi
valence, we establish a new paradigm of bit allocation using SAVI. Different fro
m previous bit allocation methods, our approach requires no empirical model and
is thus optimal. Moreover, as the original SAVI using gradient ascent only appli
es to single-level latent, we extend the SAVI to multi-level such as NVC by recu
rsively applying back-propagating through gradient ascent. Finally, we propose a
 tractable approximation for practical implementation. Our method can be applied
 to scenarios where performance outweights encoding speed, and serves as an empi
rical bound on the R-D performance of bit allocation. Experimental results show
that current state-of-the-art bit allocation algorithms still have a room of $\a
pprox 0.5$ dB PSNR to improve compared with ours. Code is available at https://g
ithub.com/tongdaxu/Bit-Allocation-Using-Optimization.
****
Regret Bounds for Markov Decision Processes with Recursive Optimized Certainty E
quivalents
Wenhao Xu, Xuefeng Gao, Xuedong He
The optimized certainty equivalent (OCE) is a family of risk measures that cover
 important examples such as entropic risk, conditional value-at-risk and mean-va
riance models. In this paper, we propose a new episodic risk-sensitive reinforce
ment learning formulation based on tabular Markov decision processes with recurs
ive OCEs. We design an efficient learning algorithm for this problem based on va
lue iteration and upper confidence bound. We derive an upper bound on the regret
 of the proposed algorithm, and also establish a minimax lower bound. Our bounds
 show that the regret rate achieved by our proposed algorithm has optimal depend
ence on the number of episodes and the number of actions.
****
Probabilistic Categorical Adversarial Attack and Adversarial Training
Han Xu, Pengfei He, Jie Ren, Yuxuan Wan, Zitao Liu, Hui Liu, Jiliang Tang
The studies on adversarial attacks and defenses have greatly improved the robust

ness of Deep Neural Networks (DNNs). Most advanced approaches have been overwhelmingly designed for continuous data such as images. However, these achievements are still hard to be generalized to categorical data. To bridge this gap, we propose a novel framework, Probabilistic Categorical Adversarial Attack (or PCAA). It transfers the discrete optimization problem of finding categorical adversarial examples to a continuous problem that can be solved via gradient-based methods. We analyze the optimality (attack success rate) and time complexity of PCAA to demonstrate its significant advantage over current search-based attacks. More importantly, through extensive empirical studies, we demonstrate that the well-established defenses for continuous data, such as adversarial training and TRADES, can be easily accommodated to defend DNNs for categorical data.
****

Hierarchical Neural Coding for Controllable CAD Model Generation
Xiang Xu, Pradeep Kumar Jayaraman, Joseph George Lambourne, Karl D.D. Willis, Yasutaka Furukawa
This paper presents a novel generative model for Computer Aided Design (CAD) that 1) represents high-level design concepts of a CAD model as a three-level hierarchical tree of neural codes, from global part arrangement down to local curve geometry; and 2) controls the generation or completion of CAD models by specifying the target design using a code tree. Concretely, a novel variant of a vector quantized VAE with "masked skip connection" extracts design variations as neural codebooks at three levels. Two-stage cascaded auto-regressive transformers learn to generate code trees from incomplete CAD models and then complete CAD models following the intended design. Extensive experiments demonstrate superior performance on conventional tasks such as unconditional generation while enabling novel interaction capabilities on conditional generation tasks. The code is available at https://github.com/samxuxiang/hnc-cad.
****

Efficient Sequence Transduction by Jointly Predicting Tokens and Durations
Hainan Xu, Fei Jia, Somshubra Majumdar, He Huang, Shinji Watanabe, Boris Ginsburg
This paper introduces a novel Token-and-Duration Transducer (TDT) architecture for sequence-to-sequence tasks. TDT extends conventional RNN-Transducer architectures by jointly predicting both a token and its duration, i.e. the number of input frames covered by the emitted token. This is achieved by using a joint network with two outputs which are independently normalized to generate distributions over tokens and durations. During inference, TDT models can skip input frames guided by the predicted duration output, which makes them significantly faster than conventional Transducers which process the encoder output frame by frame. TDT models achieve both better accuracy and significantly faster inference than conventional Transducers on different sequence transduction tasks. TDT models for Speech Recognition achieve better accuracy and up to 2.82X faster inference than conventional Transducers. TDT models for Speech Translation achieve an absolute gain of over 1 BLEU on the MUST-C test compared with conventional Transducers, and its inference is 2.27X faster. In Speech Intent Classification and Slot Filling tasks, TDT models improve the intent accuracy by up to over 1% (absolute) over conventional Transducers, while running up to 1.28X faster. Our implementation of the TDT model will be open-sourced with the NeMo (https://github.com/NVIDIA/NeMo) toolkit.
****

Constrained Efficient Global Optimization of Expensive Black-box Functions
Wenjie Xu, Yuning Jiang, Bratislav Svetozarevic, Colin Jones
We study the problem of constrained efficient global optimization, where both the objective and constraints are expensive black-box functions that can be learned with Gaussian processes. We propose CONFIG (CONstrained efFIcient Global Optimization), a simple and effective algorithm to solve it. Under certain regularity assumptions, we show that our algorithm enjoys the same cumulative regret bound as that in the unconstrained case and similar cumulative constraint violation upper bounds. For commonly used Matern and Squared Exponential kernels, our bounds are sublinear and allow us to derive a convergence rate to the optimal solutio

n of the original constrained problem. In addition, our method naturally provide s a scheme to declare infeasibility when the original black-box optimization pro blem is infeasible. Numerical experiments on sampled instances from the Gaussian process, artificial numerical problems, and a black-box building controller tun ing problem all demonstrate the competitive performance of our algorithm. Compar ed to the other state-of-the-art methods, our algorithm significantly improves t he theoretical guarantees while achieving competitive empirical performance.
****

Pareto Regret Analyses in Multi-objective Multi-armed Bandit
Mengfan Xu, Diego Klabjan
We study Pareto optimality in multi-objective multi-armed bandit by providing a formulation of adversarial multi-objective multi-armed bandit and defining its P areto regrets that can be applied to both stochastic and adversarial settings. T he regrets do not rely on any scalarization functions and reflect Pareto optimal ity compared to scalarized regrets. We also present new algorithms assuming both with and without prior information of the multi-objective multi-armed bandit se tting. The algorithms are shown optimal in adversarial settings and nearly optim al up to a logarithmic factor in stochastic settings simultaneously by our estab lished upper bounds and lower bounds on Pareto regrets. Moreover, the lower boun d analyses show that the new regrets are consistent with the existing Pareto reg ret for stochastic settings and extend an adversarial attack mechanism from band it to the multi-objective one.
****

Diverse and Faithful Knowledge-Grounded Dialogue Generation via Sequential Poste rior Inference
Yan Xu, Deqian Kong, Dehong Xu, Ziwei Ji, Bo Pang, Pascale Fung, Ying Nian Wu
The capability to generate responses with diversity and faithfulness using factu al knowledge is paramount for creating a human-like, trustworthy dialogue system . Common strategies either adopt a two-step paradigm, which optimizes knowledge selection and response generation separately, and may overlook the inherent corr elation between these two tasks, or leverage conditional variational method to j ointly optimize knowledge selection and response generation by employing an infe rence network. In this paper, we present an end-to-end learning framework, terme d Sequential Posterior Inference (SPI), capable of selecting knowledge and gener ating dialogues by approximately sampling from the posterior distribution. Unlik e other methods, SPI does not require the inference network or assume a simple g eometry of the posterior distribution. This straightforward and intuitive infere nce procedure of SPI directly queries the response generation model, allowing fo r accurate knowledge selection and generation of faithful responses. In addition to modeling contributions, our experimental results on two common dialogue data sets (Wizard of Wikipedia and Holl-E) demonstrate that SPI outperforms previous strong baselines according to both automatic and human evaluation metrics.
****

Quantifying the Variability Collapse of Neural Networks
Jing Xu, Haoxiong Liu
Recent studies empirically demonstrate the positive relationship between the tra nsferability of neural networks and the in-class variation of the last layer fea tures. The recently discovered Neural Collapse (NC) phenomenon provides a new pe rspective of understanding such last layer geometry of neural networks. In this paper, we propose a novel metric, named Variability Collapse Index (VCI), to qua ntify the variability collapse phenomenon in the NC paradigm. The VCI metric is well-motivated and intrinsically related to the linear probing loss on the last layer features. Moreover, it enjoys desired theoretical and empirical properties , including invariance under invertible linear transformations and numerical sta bility, that distinguishes it from previous metrics. Our experiments verify that VCI is indicative of the variability collapse and the transferability of pretra ined neural networks.
****

Progressive Purification for Instance-Dependent Partial Label Learning
Ning Xu, Biao Liu, Jiaqi Lv, Congyu Qiao, Xin Geng

Partial label learning (PLL) aims to train multiclass classifiers from the examp les each annotated with a set of candidate labels where a fixed but unknown cand idate label is correct. In the last few years, the instance-independent generati on process of candidate labels has been extensively studied, on the basis of whi ch many theoretical advances have been made in PLL. Nevertheless, the candidate labels are always instance-dependent in practice and there is no theoretical gua rantee that the model trained on the instance-dependent PLL examples can converg e to an ideal one. In this paper, a theoretically grounded and practically effec tive approach named POP, i.e. PrOgressive Purification for instance-dependent pa rtial label learning, is proposed. Specifically, POP updates the learning model and purifies each candidate label set progressively in every epoch. Theoreticall y, we prove that POP enlarges the region appropriately fast where the model is r eliable, and eventually approximates the Bayes optimal classifier with mild assu mptions. Technically, POP is flexible with arbitrary PLL losses and could improv e the performance of the previous PLL losses in the instance-dependent case. Exp eriments on the benchmark datasets and the real-world datasets validate the effe ctiveness of the proposed method.
****

PFGM++: Unlocking the Potential of Physics-Inspired Generative Models
Yilun Xu, Ziming Liu, Yonglong Tian, Shangyuan Tong, Max Tegmark, Tommi Jaakkola
We introduce a new family of physics-inspired generative models termed PFGM++ th at unifies diffusion models and Poisson Flow Generative Models (PFGM). These mod els realize generative trajectories for N dimensional data by embedding paths in N+D dimensional space while still controlling the progression with a simple sca lar norm of the D additional variables. The new models reduce to PFGM when D=1 a nd to diffusion models when D$\to\infty$. The flexibility of choosing D allows u s to trade off robustness against rigidity as increasing D results in more conce ntrated coupling between the data and the additional variable norms. We dispense with the biased large batch field targets used in PFGM and instead provide an u nbiased perturbation-based objective similar to diffusion models. To explore dif ferent choices of D, we provide a direct alignment method for transferring well- tuned hyperparameters from diffusion models (D$\to\infty$) to any finite D value s. Our experiments show that models with finite D can be superior to previous st ate-of-the-art diffusion models on CIFAR-10/FFHQ 64$\times$64 datasets/LSUN Chur ches 256$\times$256, with median Ds. In class-conditional setting, D=2048 yields current state-of-the-art FID of 1.74 on CIFAR-10 without additional training. F urthermore, we demonstrate that models with smaller D exhibit improved robustn ess against modeling errors. Code is available at https://github.com/Newbeeer/pf gmpp
****

Geometric Latent Diffusion Models for 3D Molecule Generation
Minkai Xu, Alexander S Powers, Ron O. Dror, Stefano Ermon, Jure Leskovec
Generative models, especially diffusion models (DMs), have achieved promising re sults for generating feature-rich geometries and advancing foundational science problems such as molecule design. Inspired by the recent huge success of Stable (latent) Diffusion models, we propose a novel and principled method for 3D molec ule generation named Geometric Latent Diffusion Models (GeoLDM). GeoLDM is the f irst latent DM model for the molecular geometry domain, composed of autoencoders encoding structures into continuous latent codes and DMs operating in the laten t space. Our key innovation is that for modeling the 3D molecular geometries, we capture its critical roto-translational equivariance constraints by building a point-structured latent space with both invariant scalars and equivariant tensor s. Extensive experiments demonstrate that GeoLDM can consistently achieve better performance on multiple molecule generation benchmarks, with up to 7% improveme nt for the valid percentage of large biomolecules. Results also demonstrate GeoL DM’s higher capacity for controllable generation thanks to the latent modeling. Code is provided at https://github.com/MinkaiXu/GeoLDM.
****

The Power of Preconditioning in Overparameterized Low-Rank Matrix Sensing
Xingyu Xu, Yandi Shen, Yuejie Chi, Cong Ma

We propose $\textsf{ScaledGD($\lambda$)}$, a preconditioned gradient descent method to tackle the low-rank matrix sensing problem when the true rank is unknown, and when the matrix is possibly ill-conditioned. Using overparametrized factor representations, $\textsf{ScaledGD($\lambda$)}$ starts from a small random initialization, and proceeds by gradient descent with a specific form of preconditioning with a fixed damping term to combat overparameterization. At the expense of light computational overhead incurred by preconditioners, $\textsf{ScaledGD($\lambda$)}$ is remarkably robust to ill-conditioning compared to vanilla gradient descent (GD). Specifically, we show that, under the Gaussian design, $\textsf{ScaledGD($\lambda$)}$ converges to the true low-rank matrix at a constant linear rate that is independent of the condition number (apart from a short nearly dimension-free burdening period), with near-optimal sample complexity. This significantly improves upon the convergence rate of vanilla GD which suffers from a polynomial dependency with the condition number. Our work provides evidence on the power of preconditioning in accelerating the convergence without hurting generalization in overparameterized learning.
****

Fascinating Supervisory Signals and Where to Find Them: Deep Anomaly Detection with Scale Learning
Hongzuo Xu, Yijie Wang, Juhui Wei, Songlei Jian, Yizhou Li, Ning Liu
Due to the unsupervised nature of anomaly detection, the key to fueling deep models is finding supervisory signals. Different from current reconstruction-guided generative models and transformation-based contrastive models, we devise novel data-driven supervision for tabular data by introducing a characteristic – scale – as data labels. By representing varied sub-vectors of data instances, we define scale as the relationship between the dimensionality of original sub-vectors and that of representations. Scales serve as labels attached to transformed representations, thus offering ample labeled data for neural network training. This paper further proposes a scale learning-based anomaly detection method. Supervised by the learning objective of scale distribution alignment, our approach learns the ranking of representations converted from varied subspaces of each data instance. Through this proxy task, our approach models inherent regularities and patterns within data, which well describes data "normality". Abnormal degrees of testing instances are obtained by measuring whether they fit these learned patterns. Extensive experiments show that our approach leads to significant improvement over state-of-the-art generative/contrastive anomaly detection methods.
****

Competing for Shareable Arms in Multi-Player Multi-Armed Bandits
Renzhe Xu, Haotian Wang, Xingxuan Zhang, Bo Li, Peng Cui
Competitions for shareable and limited resources have long been studied with strategic agents. In reality, agents often have to learn and maximize the rewards of the resources at the same time. To design an individualized competing policy, we model the competition between agents in a novel multi-player multi-armed bandit (MPMAB) setting where players are selfish and aim to maximize their own rewards. In addition, when several players pull the same arm, we assume that these players averagely share the arms' rewards by expectation. Under this setting, we first analyze the Nash equilibrium when arms' rewards are known. Subsequently, we propose a novel Selfish MPMAB with Averaging Allocation (SMAA) approach based on the equilibrium. We theoretically demonstrate that SMAA could achieve a good regret guarantee for each player when all players follow the algorithm. Additionally, we establish that no single selfish player can significantly increase their rewards through deviation, nor can they detrimentally affect other players' rewards without incurring substantial losses for themselves. We finally validate the effectiveness of the method in extensive synthetic experiments.
****

Sequential Predictive Conformal Inference for Time Series
Chen Xu, Yao Xie
We present a new distribution-free conformal prediction algorithm for sequential data (e.g., time series), called the sequential predictive conformal inference (SPCI). We specifically account for the nature that time series data are non-exc

hangeable, and thus many existing conformal prediction algorithms are not applic able. The main idea is to adaptively re-estimate the conditional quantile of non -conformity scores (e.g., prediction residuals), upon exploiting the temporal de pendence among them. More precisely, we cast the problem of conformal prediction interval as predicting the quantile of a future residual, given a user-specifie d point prediction algorithm. Theoretically, we establish asymptotic valid condi tional coverage upon extending consistency analyses in quantile regression. Usin g simulation and real-data experiments, we demonstrate a significant reduction i n interval width of SPCI compared to other existing methods under the desired em pirical coverage.
****

mPLUG-2: A Modularized Multi-modal Foundation Model Across Text, Image and Video
Haiyang Xu, Qinghao Ye, Ming Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qi Qian, Wei Wang, Guohai Xu, Ji Zhang, Songfang Huang, Fei Huang, Jing ren Zhou
Recent years have witnessed a big convergence of language, vision, and multi-mod al pretraining. In this work, we present mPLUG-2, a new unified paradigm with mo dularized design for multi-modal pretraining, which can benefit from modality co llaboration while addressing the problem of modality entanglement. In contrast t o predominant paradigms of solely relying on sequence-to-sequence generation or encoder-based instance discrimination, mPLUG-2 introduces a multi-module composi tion network by sharing common universal modules for modality collaboration and disentangling different modality modules to deal with modality entanglement. It is flexible to select different modules for different understanding and generati on tasks across all modalities including text, image, and video. Empirical study shows that mPLUG-2 achieves state-of-the-art or competitive results on a broad range of over 30 downstream tasks, spanning multi-modal tasks of image-text and video-text understanding and generation, and uni-modal tasks of text-only, image -only, and video-only understanding. Notably, mPLUG-2 shows new state-of-the-art results of 48.0 top-1 accuracy and 80.3 CIDEr on the challenging MSRVTT video Q A and video caption tasks with a far smaller model size and data scale. It also demonstrates strong zero-shot transferability on vision-language and video-langu age tasks. Code and models will be released in https://github.com/X-PLUG/mPLUG-2 .
****

ProtST: Multi-Modality Learning of Protein Sequences and Biomedical Texts
Minghao Xu, Xinyu Yuan, Santiago Miret, Jian Tang
Current protein language models (PLMs) learn protein representations mainly base d on their sequences, thereby well capturing co-evolutionary information, but th ey are unable to explicitly acquire protein functions, which is the end goal of protein representation learning. Fortunately, for many proteins, their textual p roperty descriptions are available, where their various functions are also descr ibed. Motivated by this fact, we first build the ProtDescribe dataset to augment protein sequences with text descriptions of their functions and other important properties. Based on this dataset, we propose the ProtST framework to enhance P rotein Sequence pre-training and understanding by biomedical Texts. During pre-t raining, we design three types of tasks, i.e., unimodal mask prediction, multimo dal representation alignment and multimodal mask prediction, to enhance a PLM wi th protein property information with different granularities and, at the same ti me, preserve the PLM’s original representation power. On downstream tasks, ProtS T enables both supervised learning and zero-shot prediction. We verify the super iority of ProtST-induced PLMs over previous ones on diverse representation learn ing benchmarks. Under the zero-shot setting, we show the effectiveness of ProtST on zero-shot protein classification, and ProtST also enables functional protein retrieval from a large-scale database without any function annotation.
****

Bayesian Design Principles for Frequentist Sequential Learning
Yunbei Xu, Assaf Zeevi
We develop a general theory to optimize the frequentist regret for sequential le arning problems, where efficient bandit and reinforcement learning algorithms ca

n be derived from unified Bayesian principles. We propose a novel optimization a pproach to create "algorithmic beliefs" at each round, and use Bayesian posterio rs to make decisions. This is the first approach to make Bayesian-type algorithm s prior-free and applicable to adversarial settings, in a generic and optimal ma nner. Moreover, the algorithms are simple and often efficient to implement. As a major application, we present a novel algorithm for multi-armed bandits that ac hieves the "best-of-all-worlds" empirical performance in the stochastic, adversa rial, and non-stationary environments. And we illustrate how these principles ca n be used in linear bandits, convex bandits, and reinforcement learning.
****
SLAMB: Accelerated Large Batch Training with Sparse Communication
Hang Xu, Wenxuan Zhang, Jiawei Fei, Yuzhe Wu, Tingwen Xie, Jun Huang, Yuchen Xie , Mohamed Elhoseiny, Panos Kalnis
Distributed training of large deep neural networks requires frequent exchange of massive data between machines, thus communication efficiency is a major concern . Existing compressed communication methods are either not compatible with large batch optimization algorithms, or do not provide sufficient speedup in large sc ale. In this paper, we combine sparsification-based gradient compression with th e layer-wise adaptive moments optimizer for large batch training (LAMB). We prop ose SLAMB, a novel communication-efficient optimizer that supports large batch s izes and scales to thousands of GPUs. SLAMB employs momentum masking, local erro r compensation, and element-wise adaptive rescaling to achieve accurate layer-wi se weight updates, which translates to fast convergence for very large batches. Our empirical results show that, compared to the state-of-the-art, SLAMB transmi ts half the amount of data in large-batch BERT pre-training, without sacrificing accuracy. Moreover, SLAMB achieves excellent scalability in large computing inf rastructures. For instance, SLAMB with 128 GPUs reduces the training time of Swi n Transformer pre-training on ImageNet to 5.35 hours, which is 2 hours faster th an the state-of-the-art. At the extreme, we trained BERT-XL (2.8B parameters) on 1,024 NVIDIA A100 GPUs, where SLAMB achieved 90% scaling efficiency.
****
Do Not Train It: A Linear Neural Architecture Search of Graph Neural Networks
Peng Xu, Lin Zhang, Xuanzhou Liu, Jiaqi Sun, Yue Zhao, Haiqin Yang, Bei Yu
Neural architecture search (NAS) for Graph neural networks (GNNs), called NAS-GN Ns, has achieved significant performance over manually designed GNN architecture s. However, these methods inherit issues from the conventional NAS methods, such as high computational cost and optimization difficulty. More importantly, previ ous NAS methods have ignored the uniqueness of GNNs, where GNNs possess expressi ve power without training. With the randomly-initialized weights, we can then se ek the optimal architecture parameters via the sparse coding objective and deriv e a novel NAS-GNNs method, namely neural architecture coding (NAC). Consequently , our NAC holds a no-update scheme on GNNs and can efficiently compute in linear time. Empirical evaluations on multiple GNN benchmark datasets demonstrate that our approach leads to state-of-the-art performance, which is up to $200\times$ faster and $18.8%$ more accurate than the strong baselines.
****
An Instrumental Variable Approach to Confounded Off-Policy Evaluation
Yang Xu, Jin Zhu, Chengchun Shi, Shikai Luo, Rui Song
Off-policy evaluation (OPE) aims to estimate the return of a target policy using some pre-collected observational data generated by a potentially different beha vior policy. In many cases, there exist unmeasured variables that confound the a ction-reward or action-next-state relationships, rendering many existing OPE app roaches ineffective. This paper develops an instrumental variable (IV)-based met hod for consistent OPE in confounded sequential decision making. Similar to sing le-stage decision making, we show that IV enables us to correctly identify the t arget policy’s value in infinite horizon settings as well. Furthermore, we propo se a number of policy value estimators and illustrate their effectiveness throug h extensive simulations and real data analysis from a world-leading short-video platform.
****

Near-Optimal Quantum Coreset Construction Algorithms for Clustering
Yecheng Xue, Xiaoyu Chen, Tongyang Li, Shaofeng H.-C. Jiang
k-Clustering in $\mathbb{R}^d$ (e.g., k-median and k-means) is a fundamental machine learning problem. While near-linear time approximation algorithms were known in the classical setting for a dataset with cardinality n, it remains open to find sublinear-time quantum algorithms. We give quantum algorithms that find coresets for k-clustering in $\mathbb{R}^d$ with $\tilde{O}(\sqrt{nk}d^{3/2})$ query complexity. Our coreset reduces the input size from n to $\mathrm{poly}(k\epsilon^{-1}d)$, so that existing α-approximation algorithms for clustering can run on top of it and yield $(1 + \epsilon)\alpha$-approximation. This eventually yields a quadratic speedup for various k-clustering approximation algorithms. We complement our algorithm with a nearly matching lower bound, that any quantum algorithm must make $\Omega(\sqrt{nk})$ queries in order to achieve even $O(1)$-approximation for k-clustering.
****

A Study on Transformer Configuration and Training Objective
Fuzhao Xue, Jianghai Chen, Aixin Sun, Xiaozhe Ren, Zangwei Zheng, Xiaoxin He, Yongming Chen, Xin Jiang, Yang You
Transformer-based models have delivered impressive results on many tasks, particularly vision and language tasks. In many model training situations, conventional configurations are often adopted. For example, we usually set the base model with hidden size (i.e. model width) to be 768 and the number of transformer layers (i.e. model depth) to be 12. In this paper, we revisit these conventional configurations by studying the the relationship between transformer configuration and training objective. We show that the optimal transformer configuration is closely related to the training objective. Specifically, compared with the simple classification objective, the masked autoencoder is effective in alleviating the over-smoothing issue in deep transformer training. Based on this finding, we propose “Bamboo”, a notion of using deeper and narrower transformer configurations, for masked autoencoder training. On ImageNet, with such a simple change in configuration, the re-designed Base-level transformer achieves 84.2% top-1 accuracy and outperforms SoTA models like MAE by $0.9%$. On language tasks, re-designed model outperforms BERT with the default setting by 1.1 points on average, on GLUE benchmark with 8 datasets.
****

LazyGNN: Large-Scale Graph Neural Networks via Lazy Propagation
Rui Xue, Haoyu Han, Mohamadali Torkamani, Jian Pei, Xiaorui Liu
Recent works have demonstrated the benefits of capturing long-distance dependency in graphs by deeper graph neural networks (GNNs). But deeper GNNs suffer from the long-lasting scalability challenge due to the neighborhood explosion problem in large-scale graphs. In this work, we propose to capture long-distance dependency in graphs by shallower models instead of deeper models, which leads to a much more efficient model, LazyGNN, for graph representation learning. Moreover, we demonstrate that LazyGNN is compatible with existing scalable approaches (such as sampling methods) for further accelerations through the development of mini-batch LazyGNN. Comprehensive experiments demonstrate its superior prediction performance and scalability on large-scale benchmarks. The implementation of LazyGNN is available at https: //github.com/RXPHD/Lazy_GNN.
****

Which Features are Learnt by Contrastive Learning? On the Role of Simplicity Bias in Class Collapse and Feature Suppression
Yihao Xue, Siddharth Joshi, Eric Gan, Pin-Yu Chen, Baharan Mirzasoleiman
Contrastive learning (CL) has emerged as a powerful technique for representation learning, with or without label supervision. However, supervised CL is prone to collapsing representations of subclasses within a class by not capturing all their features, and unsupervised CL may suppress harder class-relevant features by focusing on learning easy class-irrelevant features; both significantly compromise representation quality. Yet, there is no theoretical understanding of class collapse or feature suppression at test time. We provide the first unified theoretically rigorous framework to determine which features are learnt by CL. Our an

alysis indicate that, perhaps surprisingly, bias of (stochastic) gradient descen t towards finding simpler solutions is a key factor in collapsing subclass repre sentations and suppressing harder class-relevant features. Moreover, we present increasing embedding dimensionality and improving the quality of data augmentati ons as two theoretically motivated solutions to feature suppression. We also pro vide the first theoretical explanation for why employing supervised and unsuperv ised CL together yields higher-quality representations, even when using commonly -used stochastic gradient methods.
****

Adaptive Computation with Elastic Input Sequence
Fuzhao Xue, Valerii Likhosherstov, Anurag Arnab, Neil Houlsby, Mostafa Dehghani, Yang You
Humans have the ability to adapt the type of information they use, the procedure they employ, and the amount of time they spend when solving problems. However, most standard neural networks have a fixed function type and computation budget regardless of the sample’s nature or difficulty. Adaptivity is a powerful paradi gm as it not only imbues practitioners with flexibility pertaining to the downst ream usage of these models but can also serve as a powerful inductive bias for s olving certain challenging classes of problems. In this work, we introduce a new approach called AdaTape, which allows for dynamic computation in neural network s through adaptive tape tokens. AdaTape utilizes an elastic input sequence by eq uipping an architecture with a dynamic read-and-write tape. Specifically, we ada ptively generate input sequences using tape tokens obtained from a tape bank whi ch can be either trainable or derived from input data. We examine the challenges and requirements to obtain dynamic sequence content and length, and propose the Adaptive Tape Reading (ATR) algorithm to achieve both goals. Through extensive experiments on image recognition tasks, we show that AdaTape can achieve better performance while maintaining the computational cost. To facilitate further rese arch, we have released code at https://github.com/google-research/scenic/tree/ma in/scenic/projects/adatape.
****

Q-learning Decision Transformer: Leveraging Dynamic Programming for Conditional Sequence Modelling in Offline RL
Taku Yamagata, Ahmed Khalil, Raul Santos-Rodriguez
Recent works have shown that tackling offline reinforcement learning (RL) with a conditional policy produces promising results. The Decision Transformer (DT) co mbines the conditional policy approach and a transformer architecture, showing c ompetitive performance against several benchmarks. However, DT lacks stitching a bility – one of the critical abilities for offline RL to learn the optimal polic y from sub-optimal trajectories. This issue becomes particularly significant whe n the offline dataset only contains sub-optimal trajectories. On the other hand, the conventional RL approaches based on Dynamic Programming (such as Q-learning) do not have the same limitation; however, they suffer from unstable learning b ehaviours, especially when they rely on function approximation in an off-policy learning setting. In this paper, we propose the Q-learning Decision Transformer (QDT) to address the shortcomings of DT by leveraging the benefits of Dynamic Pr ogramming (Q-learning). It utilises the Dynamic Programming results to relabel t he return-to-go in the training data to then train the DT with the relabelled da ta. Our approach efficiently exploits the benefits of these two approaches and c ompensates for each other’s shortcomings to achieve better performance.
****

Quantum Ridgelet Transform: Winning Lottery Ticket of Neural Networks with Quant um Computation
Hayata Yamasaki, Sathyawageeswar Subramanian, Satoshi Hayakawa, Sho Sonoda
A significant challenge in the field of quantum machine learning (QML) is to est ablish applications of quantum computation to accelerate common tasks in machine learning such as those for neural networks. Ridgelet transform has been a funda mental mathematical tool in the theoretical studies of neural networks, but the practical applicability of ridgelet transform to conducting learning tasks was l imited since its numerical implementation by conventional classical computation

requires an exponential runtime $\exp(O(D))$ as data dimension D increases. To address this problem, we develop a quantum ridgelet transform (QRT), which implements the ridgelet transform of a quantum state within a linear runtime $O(D)$ of quantum computation. As an application, we also show that one can use QRT as a fundamental subroutine for QML to efficiently find a sparse trainable subnetwork of large shallow wide neural networks without conducting large-scale optimization of the original network. This application discovers an efficient way in this regime to demonstrate the lottery ticket hypothesis on finding such a sparse trainable neural network. These results open an avenue of QML for accelerating learning tasks with commonly used classical neural networks.
****

Compressed Decentralized Proximal Stochastic Gradient Method for Nonconvex Composite Problems with Heterogeneous Data
Yonggui Yan, Jie Chen, Pin-Yu Chen, Xiaodong Cui, Songtao Lu, Yangyang Xu
We first propose a decentralized proximal stochastic gradient tracking method (DProxSGT) for nonconvex stochastic composite problems, with data heterogeneously distributed on multiple workers in a decentralized connected network. To save communication cost, we then extend DProxSGT to a compressed method by compressing the communicated information. Both methods need only $\mathcal{O}(1)$ samples per worker for each proximal update, which is important to achieve good generalization performance on training deep neural networks. With a smoothness condition on the expected loss function (but not on each sample function), the proposed methods can achieve an optimal sample complexity result to produce a near-stationary point. Numerical experiments on training neural networks demonstrate the significantly better generalization performance of our methods over large-batch training methods and momentum variance-reduction methods and also, the ability of handling heterogeneous data by the gradient tracking scheme.
****

Temporally Consistent Transformers for Video Generation
Wilson Yan, Danijar Hafner, Stephen James, Pieter Abbeel
To generate accurate videos, algorithms have to understand the spatial and temporal dependencies in the world. Current algorithms enable accurate predictions over short horizons but tend to suffer from temporal inconsistencies. When generated content goes out of view and is later revisited, the model invents different content instead. Despite this severe limitation, no established benchmarks exist for video generation with long temporal dependencies. In this paper, we curate 3 challenging video datasets with long-range dependencies by rendering walks through 3D scenes of procedural mazes, Minecraft worlds, and indoor scans. We perform a comprehensive evaluation of current models and observe their limitations in temporal consistency. Moreover, we introduce the Temporally Consistent Transformer (TECO), a generative model that substantially improves long-term consistency while also reducing sampling time. By compressing its input sequence into fewer embeddings, applying a temporal transformer, and expanding back using a spatial MaskGit, TECO outperforms existing models across many metrics. Videos are available on the website: https://wilson1yan.github.io/teco
****

Distortion and Uncertainty Aware Loss for Panoramic Depth Completion
Zhiqiang Yan, Xiang Li, Kun Wang, Shuo Chen, Jun Li, Jian Yang
Standard MSE or MAE loss function is commonly used in limited field-of-vision depth completion, treating each pixel equally under a basic assumption that all pixels have same contribution during optimization. Recently, with the rapid rise of panoramic photography, panoramic depth completion (PDC) has raised increasing attention in 3D computer vision. However, the assumption is inapplicable to panoramic data due to its latitude-wise distortion and high uncertainty nearby textures and edges. To handle these challenges, we propose distortion and uncertainty aware loss (DUL) that consists of a distortion-aware loss and an uncertainty-aware loss. The distortion-aware loss is designed to tackle the panoramic distortion caused by equirectangular projection, whose coordinate transformation relation is used to adaptively calculate the weight of the latitude-wise distortion, distributing uneven importance instead of the equal treatment for each pixel. The

uncertainty-aware loss is presented to handle the inaccuracy in non-smooth regions. Specifically, we characterize uncertainty into PDC solutions under Bayesian deep learning framework, where a novel consistent uncertainty estimation constraint is designed to learn the consistency between multiple uncertainty maps of a single panorama. This consistency constraint allows model to produce more precise uncertainty estimation that is robust to feature deformation. Extensive experiments show the superiority of our method over standard loss functions, reaching the state of the art.

****

Self-Interpretable Time Series Prediction with Counterfactual Explanations

Jingquan Yan, Hao Wang

Interpretable time series prediction is crucial for safety-critical areas such as healthcare and autonomous driving. Most existing methods focus on interpreting predictions by assigning important scores to segments of time series. In this paper, we take a different and more challenging route and aim at developing a self-interpretable model, dubbed Counterfactual Time Series (CounTS), which generates counterfactual and actionable explanations for time series predictions. Specifically, we formalize the problem of time series counterfactual explanations, establish associated evaluation protocols, and propose a variational Bayesian deep learning model equipped with counterfactual inference capability of time series abduction, action, and prediction. Compared with state-of-the-art baselines, our self-interpretable model can generate better counterfactual explanations while maintaining comparable prediction accuracy.

****

Quantum 3D Graph Learning with Applications to Molecule Embedding

Ge Yan, Huaijin Wu, Junchi Yan

Learning 3D graph with spatial position as well as node attributes has been recently actively studied, for its utility in different applications e.g. 3D molecules. Quantum computing is known a promising direction for its potential theoretical supremacy for large-scale graph and combinatorial problem as well as the increasing evidence for the availability to physical quantum devices in the near term. In this paper, for the first time to our best knowledge, we propose a quantum 3D embedding ansatz that learns the latent representation of 3D structures from the Hilbert space composed of the Bloch sphere of each qubit. Specifically, the 3D Cartesian coordinates of nodes are converted into rotation and torsion angles and then encode them into the form of qubits. Moreover, Parameterized Quantum Circuit (PQC) is applied to serve as the trainable layers and the output of the PQC is adopted as the final node embedding. Experimental results on two downstream tasks, molecular property prediction and 3D molecular geometries generation, demonstrate the effectiveness of our model. We show the capacity and capability of our model with the evaluation on the QM9 dataset (134k molecules) with very few parameters, and its potential to be executed on a real quantum device.

****

Fast Rates in Time-Varying Strongly Monotone Games

Yu-Hu Yan, Peng Zhao, Zhi-Hua Zhou

Multi-player online games depict the interaction of multiple players with each other over time. Strongly monotone games are of particular interest since they have benign properties and also relate to many classic games that have applications in real life. Existing works mainly focus on the time-invariant case with provable guarantees established. However, the research of the more general time-varying games in changing environments is underexplored and the best-known result cannot match the guarantees in the time-invariant case. In this work, we present a new decentralized online algorithm for time-varying strongly monotone games, which greatly improves existing results and obtains fast rates, matching the best time-invariant guarantee without knowing the environmental non-stationarity. Furthermore, to achieve faster rates, we generalize the RVU property with smoothness and establish a series of problem-dependent bounds that also match the best time-invariant one. To realize all those results, we make a comprehensive use of the techniques in non-stationary and universal online learning.

****

Proper Scoring Rules for Survival Analysis
Hiroki Yanagisawa
Survival analysis is the problem of estimating probability distributions for future event times, which can be seen as a problem in uncertainty quantification. Although there are fundamental theories on strictly proper scoring rules for uncertainty quantification, little is known about those for survival analysis. In this paper, we investigate extensions of four major strictly proper scoring rules for survival analysis and we prove that these extensions are proper under certain conditions, which arise from the discretization of the estimation of probability distributions. We also compare the estimation performances of these extended scoring rules by using real datasets, and the extensions of the logarithmic score and the Brier score performed the best.
****
Behavior Contrastive Learning for Unsupervised Skill Discovery
Rushuai Yang, Chenjia Bai, Hongyi Guo, Siyuan Li, Bin Zhao, Zhen Wang, Peng Liu, Xuelong Li
In reinforcement learning, unsupervised skill discovery aims to learn diverse skills without extrinsic rewards. Previous methods discover skills by maximizing the mutual information (MI) between states and skills. However, such an MI objective tends to learn simple and static skills and may hinder exploration. In this paper, we propose a novel unsupervised skill discovery method through contrastive learning among behaviors, which makes the agent produce similar behaviors for the same skill and diverse behaviors for different skills. Under mild assumptions, our objective maximizes the MI between different behaviors based on the same skill, which serves as an upper bound of the previous MI objective. Meanwhile, our method implicitly increases the state entropy to obtain better state coverage. We evaluate our method on challenging mazes and continuous control tasks. The results show that our method generates diverse and far-reaching skills, and also obtains competitive performance in downstream tasks compared to the state-of-the-art methods.
****
Nested Elimination: A Simple Algorithm for Best-Item Identification From Choice-Based Feedback
Junwen Yang, Yifan Feng
We study the problem of best-item identification from choice-based feedback. In this problem, a company sequentially and adaptively shows display sets to a population of customers and collects their choices. The objective is to identify the most preferred item with the least number of samples and at a high confidence level. We propose an elimination-based algorithm, namely Nested Elimination (NE), which is inspired by the nested structure implied by the information-theoretic lower bound. NE is simple in structure, easy to implement, and has a strong theoretical guarantee for sample complexity. Specifically, NE utilizes an innovative elimination criterion and circumvents the need to solve any complex combinatorial optimization problem. We provide an instance-specific and non-asymptotic bound on the expected sample complexity of NE. We also show NE achieves high-order worst-case asymptotic optimality. Finally, numerical experiments from both synthetic and real data corroborate our theoretical findings.
****
Towards Better Graph Representation Learning with Parameterized Decomposition & Filtering
Mingqi Yang, Wenjie Feng, Yanming Shen, Bryan Hooi
Proposing an effective and flexible matrix to represent a graph is a fundamental challenge that has been explored from multiple perspectives, e.g., filtering in Graph Fourier Transforms. In this work, we develop a novel and general framework which unifies many existing GNN models from the view of parameterized decomposition and filtering, and show how it helps to enhance the flexibility of GNNs while alleviating the smoothness and amplification issues of existing models. Essentially, we show that the extensively studied spectral graph convolutions with learnable polynomial filters are constrained variants of this formulation, and releasing these constraints enables our model to express the desired decomposition

and filtering simultaneously. Based on this generalized framework, we develop models that are simple in implementation but achieve significant improvements and computational efficiency on a variety of graph learning tasks. Code is available at https://github.com/qslim/PDF.

****

Weighted Flow Diffusion for Local Graph Clustering with Node Attributes: an Algorithm and Statistical Guarantees

Shenghao Yang, Kimon Fountoulakis

Local graph clustering methods aim to detect small clusters in very large graphs without the need to process the whole graph. They are fundamental and scalable tools for a wide range of tasks such as local community detection, node ranking and node embedding. While prior work on local graph clustering mainly focuses on graphs without node attributes, modern real-world graph datasets typically come with node attributes that provide valuable additional information. We present a simple local graph clustering algorithm for graphs with node attributes, based on the idea of diffusing mass locally in the graph while accounting for both structural and attribute proximities. Using high-dimensional concentration results, we provide statistical guarantees on the performance of the algorithm for the recovery of a target cluster with a single seed node. We give conditions under which a target cluster generated from a fairly general contextual random graph model, which includes both the stochastic block model and the planted cluster model as special cases, can be fully recovered with bounded false positives. Empirically, we validate all theoretical claims using synthetic data, and we show that incorporating node attributes leads to superior local clustering performances using real-world graph datasets.

****

Chemically Transferable Generative Backmapping of Coarse-Grained Proteins

Soojung Yang, Rafael Gomez-Bombarelli

Coarse-graining (CG) accelerates molecular simulations of protein dynamics by simulating sets of atoms as singular beads. Backmapping is the opposite operation of bringing lost atomistic details back from the CG representation. While machine learning (ML) has produced accurate and efficient CG simulations of proteins, fast and reliable backmapping remains a challenge. Rule-based methods produce poor all-atom geometries, needing computationally costly refinement through additional simulations. Recently proposed ML approaches outperform traditional baselines but are not transferable between proteins and sometimes generate unphysical atom placements with steric clashes and implausible torsion angles. This work addresses both issues to build a fast, transferable, and reliable generative backmapping tool for CG protein representations. We achieve generalization and reliability through a combined set of innovations: representation based on internal coordinates; an equivariant encoder/prior; a custom loss function that helps ensure local structure, global structure, and physical constraints; and expert curation of high-quality out-of-equilibrium protein data for training. Our results pave the way for out-of-the-box backmapping of coarse-grained simulations for arbitrary proteins.

****

Data Poisoning Attacks Against Multimodal Encoders

Ziqing Yang, Xinlei He, Zheng Li, Michael Backes, Mathias Humbert, Pascal Berrang, Yang Zhang

Recently, the newly emerged multimodal models, which leverage both visual and linguistic modalities to train powerful encoders, have gained increasing attention. However, learning from a large-scale unlabeled dataset also exposes the model to the risk of potential poisoning attacks, whereby the adversary aims to perturb the model’s training data to trigger malicious behaviors in it. In contrast to previous work, only poisoning visual modality, in this work, we take the first step to studying poisoning attacks against multimodal models in both visual and linguistic modalities. Specially, we focus on answering two questions: (1) Is the linguistic modality also vulnerable to poisoning attacks? and (2) Which modality is most vulnerable? To answer the two questions, we propose three types of poisoning attacks against multimodal models. Extensive evaluations on different da

tasets and model architectures show that all three attacks can achieve significa nt attack performance while maintaining model utility in both visual and linguis tic modalities. Furthermore, we observe that the poisoning effect differs betwee n different modalities. To mitigate the attacks, we propose both pre-training an d post-training defenses. We empirically show that both defenses can significant ly reduce the attack performance while preserving the model's utility. Our code is available at https://github.com/zqypku/mm_poison/.

****

Towards Sustainable Learning: Coresets for Data-efficient Deep Learning

Yu Yang, Hao Kang, Baharan Mirzasoleiman

To improve the efficiency and sustainability of learning deep models, we propose CREST, the first scalable framework with rigorous theoretical guarantees to ide ntify the most valuable examples for training non-convex models, particularly de ep networks. To guarantee convergence to a stationary point of a non-convex func tion, CREST models the non-convex loss as a series of quadratic functions and ex tracts a coreset for each quadratic sub-region. In addition, to ensure faster co nvergence of stochastic gradient methods such as (mini-batch) SGD, CREST iterati vely extracts multiple mini-batch coresets from larger random subsets of trainin g data, to ensure nearly-unbiased gradients with small variances. Finally, to fu rther improve scalability and efficiency, CREST identifies and excludes the exam ples that are learned from the coreset selection pipeline. Our extensive experim ents on several deep networks trained on vision and NLP datasets, including CIFA R-10, CIFAR-100, TinyImageNet, and SNLI, confirm that CREST speeds up training d eep networks on very large datasets, by 1.7x to 2.5x with minimum loss in the pe rformance. By analyzing the learning difficulty of the subsets selected by CREST , we show that deep models benefit the most by learning from subsets of increasi ng difficulty levels.

****

Improving Adversarial Robustness by Putting More Regularizations on Less Robust Samples

Dongyoon Yang, Insung Kong, Yongdai Kim

Adversarial training, which is to enhance robustness against adversarial attacks , has received much attention because it is easy to generate human-imperceptible perturbations of data to deceive a given deep neural network. In this paper, we propose a new adversarial training algorithm that is theoretically well motivat ed and empirically superior to other existing algorithms. A novel feature of the proposed algorithm is to apply more regularization to data vulnerable to advers arial attacks than other existing regularization algorithms do. Theoretically, w e show that our algorithm can be understood as an algorithm of minimizing a newl y derived upper bound of the robust risk. Numerical experiments illustrate that our proposed algorithm improves the generalization (accuracy on examples) and ro bustness (accuracy on adversarial attacks) simultaneously to achieve the state-o f-the-art performance.

****

Improving Adversarial Robustness of Deep Equilibrium Models with Explicit Regula tions Along the Neural Dynamics

Zonghan Yang, Peng Li, Tianyu Pang, Yang Liu

Deep equilibrium (DEQ) models replace the multiple-layer stacking of conventiona l deep networks with a fixed-point iteration of a single-layer transformation. H aving been demonstrated to be competitive in a variety of real-world scenarios, the adversarial robustness of general DEQs becomes increasingly crucial for thei r reliable deployment. Existing works improve the robustness of general DEQ mode ls with the widely-used adversarial training (AT) framework, but they fail to ex ploit the structural uniquenesses of DEQ models. To this end, we interpret DEQs through the lens of neural dynamics and find that AT under-regulates intermediat e states. Besides, the intermediate states typically provide predictions with a high prediction entropy. Informed by the correlation between the entropy of dyna mical systems and their stability properties, we propose reducing prediction ent ropy by progressively updating inputs along the neural dynamics. During AT, we a lso utilize random intermediate states to compute the loss function. Our methods

regulate the neural dynamics of DEQ models in this manner. Extensive experiments demonstrate that our methods substantially increase the robustness of DEQ models and even outperform the strong deep network baselines.

****

Mitigating Spurious Correlations in Multi-modal Models during Fine-tuning

Yu Yang, Besmira Nushi, Hamid Palangi, Baharan Mirzasoleiman

Spurious correlations that degrade model generalization or lead the model to be right for the wrong reasons are one of the main robustness concerns for real-world deployments. However, mitigating these correlations during pre-training for large-scale models can be costly and impractical, particularly for those without access to high-performance computing resources. This paper proposes a novel approach to address spurious correlations during fine-tuning for a given domain of interest. With a focus on multi-modal models (e.g., CLIP), the proposed method leverages different modalities in these models to detect and explicitly set apart spurious attributes from the affected class, achieved through a multi-modal contrastive loss function that expresses spurious relationships through language. Our experimental results and in-depth visualizations on CLIP show that such an intervention can effectively i) improve the model’s accuracy when spurious attributes are not present, and ii) directs the model’s activation maps towards the actual class rather than the spurious attribute when present. In particular, on the Waterbirds dataset, our algorithm achieved a worst-group accuracy 23% higher than ERM on CLIP with a ResNet-50 backbone, and 32% higher on CLIP with a ViT backbone, while maintaining the same average accuracy as ERM.

****

A theory of representation learning gives a deep generalisation of kernel methods

Adam X. Yang, Maxime Robeyns, Edward Milsom, Ben Anson, Nandi Schoots, Laurence Aitchison

The successes of modern deep machine learning methods are founded on their ability to transform inputs across multiple layers to build good high-level representations. It is therefore critical to understand this process of representation learning. However, standard theoretical approaches (formally NNGPs) involving infinite width limits eliminate representation learning. We therefore develop a new infinite width limit, the Bayesian representation learning limit, that exhibits representation learning mirroring that in finite-width models, yet at the same time, retains some of the simplicity of standard infinite-width limits. In particular, we show that Deep Gaussian processes (DGPs) in the Bayesian representation learning limit have exactly multivariate Gaussian posteriors, and the posterior covariances can be obtained by optimizing an interpretable objective combining a log-likelihood to improve performance with a series of KL-divergences which keep the posteriors close to the prior. We confirm these results experimentally in wide but finite DGPs. Next, we introduce the possibility of using this limit and objective as a flexible, deep generalisation of kernel methods, that we call deep kernel machines (DKMs). Like most naive kernel methods, DKMs scale cubically in the number of datapoints. We therefore use methods from the Gaussian process inducing point literature to develop a sparse DKM that scales linearly in the number of datapoints. Finally, we extend these approaches to NNs (which have non-Gaussian posteriors) in the Appendices.

****

Efficient Algorithms for Exact Graph Matching on Correlated Stochastic Block Models with Constant Correlation

Joonhyuk Yang, Dongpil Shin, Hye Won Chung

We consider the problem of graph matching, or learning vertex correspondence, between two correlated stochastic block models (SBMs). The graph matching problem arises in various fields, including computer vision, natural language processing and bioinformatics, and in particular, matching graphs with inherent community structure has significance related to de-anonymization of correlated social networks. Compared to the correlated Erdos-Renyi (ER) model, where various efficient algorithms have been developed, among which a few algorithms have been proven to achieve the exact matching with constant edge correlation, no low-order polyno

mial algorithm has been known to achieve exact matching for the correlated SBMs with constant correlation. In this work, we propose an efficient algorithm for matching graphs with community structure, based on the comparison between partition trees rooted from each vertex, by extending the idea of Mao et al. (2021) to graphs with communities. The partition tree divides the large neighborhoods of each vertex into disjoint subsets using their edge statistics to different communities. Our algorithm is the first low-order polynomial-time algorithm achieving exact matching between two correlated SBMs with high probability in dense graphs.

****

Are Neurons Actually Collapsed? On the Fine-Grained Structure in Neural Representations

Yongyi Yang, Jacob Steinhardt, Wei Hu

Recent work has observed an intriguing "Neural Collapse" phenomenon in well-trained neural networks, where the last-layer representations of training samples with the same label collapse into each other. This appears to suggest that the last-layer representations are completely determined by the labels, and do not depend on the intrinsic structure of input distribution. We provide evidence that this is not a complete description, and that the apparent collapse hides important fine-grained structure in the representations. Specifically, even when representations apparently collapse, the small amount of remaining variation can still faithfully and accurately captures the intrinsic structure of input distribution. As an example, if we train on CIFAR-10 using only 5 coarse-grained labels (by combining two classes into one super-class) until convergence, we can reconstruct the original 10-class labels from the learned representations via unsupervised clustering. The reconstructed labels achieve 93% accuracy on the CIFAR-10 test set, nearly matching the normal CIFAR-10 accuracy for the same architecture. We also provide an initial theoretical result showing the fine-grained representation structure in a simplified synthetic setting. Our results show concretely how the structure of input data can play a significant role in determining the fine-grained structure of neural representations, going beyond what Neural Collapse predicts.

****

Generative Adversarial Symmetry Discovery

Jianke Yang, Robin Walters, Nima Dehmamy, Rose Yu

Despite the success of equivariant neural networks in scientific applications, they require knowing the symmetry group a priori. However, it may be difficult to know which symmetry to use as an inductive bias in practice. Enforcing the wrong symmetry could even hurt the performance. In this paper, we propose a framework, LieGAN, to automatically discover equivariances from a dataset using a paradigm akin to generative adversarial training. Specifically, a generator learns a group of transformations applied to the data, which preserve the original distribution and fool the discriminator. LieGAN represents symmetry as interpretable Lie algebra basis and can discover various symmetries such as the rotation group $\mathrm{SO}(n)$, restricted Lorentz group $\mathrm{SO}(1,3)^+$ in trajectory prediction and top-quark tagging tasks. The learned symmetry can also be readily used in several existing equivariant neural networks to improve accuracy and generalization in prediction.

****

Boosting Offline Reinforcement Learning with Action Preference Query

Qisen Yang, Shenzhi Wang, Matthieu Gaetan Lin, Shiji Song, Gao Huang

Training practical agents usually involve offline and online reinforcement learning (RL) to balance the policy's performance and interaction costs. In particular, online fine-tuning has become a commonly used method to correct the erroneous estimates of out-of-distribution data learned in the offline training phase. However, even limited online interactions can be inaccessible or catastrophic for high-stake scenarios like healthcare and autonomous driving. In this work, we introduce an interaction-free training scheme dubbed Offline-with-Action-Preferences (OAP). The main insight is that, compared to online fine-tuning, querying the preferences between pre-collected and learned actions can be equally or even mo

re helpful to the erroneous estimate problem. By adaptively encouraging or suppr
essing policy constraint according to action preferences, OAP could distinguish
overestimation from beneficial policy improvement and thus attains a more accura
te evaluation of unseen data. Theoretically, we prove a lower bound of the behav
ior policy’s performance improvement brought by OAP. Moreover, comprehensive exp
eriments on the D4RL benchmark and state-of-the-art algorithms demonstrate that
OAP yields higher (29% on average) scores, especially on challenging AntMaze tas
ks (98% higher).
****

Towards Controlled Data Augmentations for Active Learning
Jianan Yang, Haobo Wang, Sai Wu, Gang Chen, Junbo Zhao
The mission of active learning is to identify the most valuable data samples, th
us attaining decent performance with much fewer samples. The data augmentation t
echniques seem straightforward yet promising to enhance active learning by exten
ding the exploration of the input space, which helps locate more valuable sample
s. In this work, we thoroughly study the coupling of data augmentation and activ
e learning, thereby proposing Controllable Augmentation ManiPulator for Active L
earning. In contrast to the few prior works that touched on this line, CAMPAL em
phasizes a purposeful, tighten, and better-controlled integration of data augmen
tation into active learning in three folds: (i)-carefully designed augmentation
policies applied separately on labeled and unlabeled data pools; (ii)-controlled
 and quantifiably optimizable augmentation strengths; (iii)-full and flexible co
verage for most (if not all) active learning schemes. Theories are proposed and
associated with the development of key components in CAMPAL. Through extensive e
mpirical experiments, we bring the performance of active learning methods to a n
ew level: an absolute performance boost of 16.99% on CIFAR-10 and 12.25 on SVHN
with 1,000 annotated samples. Codes are available at https://github.com/jnzju/CA
MPAL.
****

What is Essential for Unseen Goal Generalization of Offline Goal-conditioned RL?
Rui Yang, Lin Yong, Xiaoteng Ma, Hao Hu, Chongjie Zhang, Tong Zhang
Offline goal-conditioned RL (GCRL) offers a way to train general-purpose agents
from fully offline datasets. In addition to being conservative within the datase
t, the generalization ability to achieve unseen goals is another fundamental cha
llenge for offline GCRL. However, to the best of our knowledge, this problem has
 not been well studied yet. In this paper, we study out-of-distribution (OOD) ge
neralization of offline GCRL both theoretically and empirically to identify fact
ors that are important. In a number of experiments, we observe that weighted imi
tation learning enjoys better generalization than pessimism-based offline RL met
hod. Based on this insight, we derive a theory for OOD generalization, which cha
racterizes several important design choices. We then propose a new offline GCRL
method, Generalizable Offline goAl-condiTioned RL (GOAT), by combining the findi
ngs from our theoretical and empirical studies. On a new benchmark containing 9
independent identically distributed (IID) tasks and 17 OOD tasks, GOAT outperfor
ms current state-of-the-art methods by a large margin.
****

Neural Prediction Errors enable Analogical Visual Reasoning in Human Standard In
telligence Tests
Lingxiao Yang, Hongzhi You, Zonglei Zhen, Dahui Wang, Xiaohong Wan, Xiaohua Xie,
 Ru-Yuan Zhang
Deep neural networks have long been criticized for lacking the ability to perfor
m analogical visual reasoning. Here, we propose a neural network model to solve
Raven’s Progressive Matrices (RPM) - one of the standard intelligence tests in h
uman psychology. Specifically, we design a reasoning block based on the well-kno
wn concept of prediction error (PE) in neuroscience. Our reasoning block uses co
nvolution to extract abstract rules from high-level visual features of the 8 con
text images and generates the features of a predicted answer. PEs are then calcu
lated between the predicted features and those of the 8 candidate answers, and a
re then passed to the next stage. We further integrate our novel reasoning block
s into a residual network and build a new Predictive Reasoning Network (PredRNet

). Extensive experiments show that our proposed PredRNet achieves state-of-the-art average performance on several important RPM benchmarks. PredRNet also shows good generalization abilities in a variety of out-of-distribution scenarios and other visual reasoning tasks. Most importantly, our PredRNet forms low-dimensional representations of abstract rules and minimizes hierarchical prediction errors during model training, supporting the critical role of PE minimization in visual reasoning. Our work highlights the potential of using neuroscience theories to solve abstract visual reasoning problems in artificial intelligence. The code is available at https://github.com/ZjjConan/AVR-PredRNet.
****

Change is Hard: A Closer Look at Subpopulation Shift
Yuzhe Yang, Haoran Zhang, Dina Katabi, Marzyeh Ghassemi
Machine learning models often perform poorly on subgroups that are underrepresented in the training data. Yet, little is understood on the variation in mechanisms that cause subpopulation shifts, and how algorithms generalize across such diverse shifts at scale. In this work, we provide a fine-grained analysis of subpopulation shift. We first propose a unified framework that dissects and explains common shifts in subgroups. We then establish a comprehensive benchmark of 20 state-of-the-art algorithms evaluated on 12 real-world datasets in vision, language, and healthcare domains. With results obtained from training over 10,000 models, we reveal intriguing observations for future progress in this space. First, existing algorithms only improve subgroup robustness over certain types of shifts but not others. Moreover, while current algorithms rely on group-annotated validation data for model selection, we find that a simple selection criterion based on worst-class accuracy is surprisingly effective even without any group information. Finally, unlike existing works that solely aim to improve worst-group accuracy (WGA), we demonstrate the fundamental tradeoff between WGA and other important metrics, highlighting the need to carefully choose testing metrics. Code and data are available at: https://github.com/YyzHarry/SubpopBench.
****

Continual Task Allocation in Meta-Policy Network via Sparse Prompting
Yijun Yang, Tianyi Zhou, Jing Jiang, Guodong Long, Yuhui Shi
How to train a generalizable meta-policy by continually learning a sequence of tasks? It is a natural human skill yet challenging to achieve by current reinforcement learning: the agent is expected to quickly adapt to new tasks (plasticity) meanwhile retaining the common knowledge from previous tasks (stability). We address it by "Continual Task Allocation via Sparse Prompting (CoTASP)", which learns over-complete dictionaries to produce sparse masks as prompts extracting a sub-network for each task from a meta-policy network. CoTASP trains a policy for each task by optimizing the prompts and the sub-network weights alternatively. The dictionary is then updated to align the optimized prompts with tasks' embedding, thereby capturing tasks' semantic correlations. Hence, relevant tasks share more neurons in the meta-policy network due to similar prompts while cross-task interference causing forgetting is effectively restrained. Given a meta-policy and dictionaries trained on previous tasks, new task adaptation reduces to highly efficient sparse prompting and sub-network finetuning. In experiments, CoTASP achieves a promising plasticity-stability trade-off without storing or replaying any past tasks' experiences. It outperforms existing continual and multi-task RL methods on all seen tasks, forgetting reduction, and generalization to unseen tasks.
****

Hyperbolic Representation Learning: Revisiting and Advancing
Menglin Yang, Min Zhou, Rex Ying, Yankai Chen, Irwin King
The non-Euclidean geometry of hyperbolic spaces has recently garnered considerable attention in the realm of representation learning. Current endeavors in hyperbolic representation largely presuppose that the underlying hierarchies can be automatically inferred and preserved through the adaptive optimization process. This assumption, however, is questionable and requires further validation. In this work, we first introduce a position-tracking mechanism to scrutinize existing prevalent hyperbolic models, revealing that the learned representations are sub-

optimal and unsatisfactory. To address this, we propose a simple yet effective m
ethod, hyperbolic informed embedding (HIE), by incorporating cost-free hierarchi
cal information deduced from the hyperbolic distance of the node to the origin (
i.e., induced hyperbolic norm) to advance existing hyperbolic models. The propos
ed method HIE is both task-agnostic and model-agnostic, enabling its seamless in
tegration with a broad spectrum of models and tasks. Extensive experiments acros
s various models and different tasks demonstrate the versatility and adaptabilit
y of the proposed method. Remarkably, our method achieves a remarkable improveme
nt of up to 21.4% compared to the competing baselines.
****

Which is Better for Learning with Noisy Labels: The Semi-supervised Method or Mo
deling Label Noise?

Yu Yao, Mingming Gong, Yuxuan Du, Jun Yu, Bo Han, Kun Zhang, Tongliang Liu

In real life, accurately annotating large-scale datasets is sometimes difficult.
 Datasets used for training deep learning models are likely to contain label noi
se. To make use of the dataset containing label noise, two typical methods have
been proposed. One is to employ the semi-supervised method by exploiting labeled
 confident examples and unlabeled unconfident examples. The other one is to mode
l label noise and design statistically consistent classifiers. A natural questio
n remains unsolved: which one should be used for a specific real-world applicati
on? In this paper, we answer the question from the perspective of causal data ge
nerative process. Specifically, the performance of the semi-supervised based met
hod depends heavily on the data generative process while the method modeling lab
el-noise is not influenced by the generation process. For example, for a given d
ataset, if it has a causal generative structure that the features cause the labe
l, the semi-supervised based method would not be helpful. When the causal struct
ure is unknown, we provide an intuitive method to discover the causal structure
for a given dataset containing label noise.
****

How Bad is Top-K Recommendation under Competing Content Creators?

Fan Yao, Chuanhao Li, Denis Nekipelov, Hongning Wang, Haifeng Xu

This study explores the impact of content creators' competition on user welfare
in recommendation platforms, as well as the long-term dynamics of relevance-driv
en recommendations. We establish a model of creator competition, under the setti
ng where the platform uses a top-K recommendation policy, user decisions are g
uided by the Random Utility model, and creators, in absence of explicit utility
functions, employ arbitrary no-regret learning algorithms for strategy updates.
We study the user welfare guarantee through the lens of Price of Anarchy and sho
w that the fraction of user welfare loss due to creator competition is always up
per bounded by a small constant depending on K and randomness in user decision
s; we also prove the tightness of this bound. Our result discloses an intrinsic
merit of the relevance-driven recommendation policy, as long as users' decisions
 involve randomness and the platform provides reasonably many alternatives to it
s users.
****

MultiAdam: Parameter-wise Scale-invariant Optimizer for Multiscale Training of P
hysics-informed Neural Networks

Jiachen Yao, Chang Su, Zhongkai Hao, Songming Liu, Hang Su, Jun Zhu

Physics-informed Neural Networks (PINNs) have recently achieved remarkable progr
ess in solving Partial Differential Equations (PDEs) in various fields by minimi
zing a weighted sum of PDE loss and boundary loss. However, there are several cr
itical challenges in the training of PINNs, including the lack of theoretical fr
ameworks and the imbalance between PDE loss and boundary loss. In this paper, we
 present an analysis of second-order non-homogeneous PDEs, which are classified
into three categories and applicable to various common problems. We also charact
erize the connections between the training loss and actual error, guaranteeing c
onvergence under mild conditions. The theoretical analysis inspires us to furthe
r propose MultiAdam, a scale-invariant optimizer that leverages gradient momentu
m to parameter-wisely balance the loss terms. Extensive experiment results on mu
ltiple problems from different physical domains demonstrate that our MultiAdam s

olver can improve the predictive accuracy by 1-2 orders of magnitude compared wi
th strong baselines.
****
Policy Mirror Ascent for Efficient and Independent Learning in Mean Field Games
Batuhan Yardim, Semih Cayci, Matthieu Geist, Niao He
Mean-field games have been used as a theoretical tool to obtain an approximate N
ash equilibrium for symmetric and anonymous N-player games. However, limiting
applicability, existing theoretical results assume variations of a “population g
enerative model”, which allows arbitrary modifications of the population distrib
ution by the learning algorithm. Moreover, learning algorithms typically work on
 abstract simulators with population instead of the N-player game. Instead, we
 show that N agents running policy mirror ascent converge to the Nash equilibr
ium of the regularized game within $\widetilde{\mathcal{O}}(\varepsilon^{-2})$ s
amples from a single sample trajectory without a population generative model, up
 to a standard $\mathcal{O}(\frac{1}{\sqrt{N}})$ error due to the mean field. Ta
king a divergent approach from the literature, instead of working with the best-
response map we first show that a policy mirror ascent map can be used to constr
uct a contractive operator having the Nash equilibrium as its fixed point. We an
alyze single-path TD learning for N-agent games, proving sample complexity gua
rantees by only using a sample path from the N-agent simulator without a popul
ation generative model. Furthermore, we demonstrate that our methodology allows
 for independent learning by N agents with finite sample guarantees.
****
Retrieval-Augmented Multimodal Language Modeling
Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Richard James, Jure Leskovec,
Percy Liang, Mike Lewis, Luke Zettlemoyer, Wen-Tau Yih
Recent multimodal models such as DALL-E and CM3 have achieved remarkable progres
s in text-to-image and image-to-text generation. However, these models store all
 their knowledge (e.g., the appearance of the Eiffel Tower) in the model paramet
ers, requiring increasingly larger models and training data to capture more know
ledge. To integrate knowledge in a more scalable and modular way, we propose a r
etrieval-augmented multimodal model, which enables a base multimodal model (gene
rator) to refer to relevant text and images fetched by a retriever from external
 memory (e.g., documents on the web). Specifically, for the retriever, we use a
pretrained CLIP, and for the generator, we train a CM3 Transformer on the LAION
dataset. Our resulting model, named Retrieval-Augmented CM3 (RA-CM3), is the fir
st multimodal model that can retrieve and generate both text and images. We show
 that RA-CM3 significantly outperforms baseline multimodal models such as DALL-E
 and CM3 on both image and caption generation tasks (12 FID and 17 CIDEr improve
ments on MS-COCO), while requiring much less compute for training ($<$30% of DAL
L-E). Moreover, we show that RA-CM3 exhibits novel capabilities such as faithful
 image generation and multimodal in-context learning (e.g., image generation fro
m demonstrations).
****
On the Power of Pre-training for Generalization in RL: Provable Benefits and Har
dness
Haotian Ye, Xiaoyu Chen, Liwei Wang, Simon Shaolei Du
Generalization in Reinforcement Learning (RL) aims to train an agent during trai
ning that generalizes to the target environment. In this work, we first point ou
t that RL generalization is fundamentally different from the generalization in s
upervised learning, and fine-tuning on the target environment is necessary for g
ood test performance. Therefore, we seek to answer the following question: how m
uch can we expect pre-training over training environments to be helpful for effi
cient and effective fine-tuning? On one hand, we give a surprising result showin
g that asymptotically, the improvement from pre-training is at most a constant f
actor. On the other hand, we show that pre-training can be indeed helpful in the
 non-asymptotic regime by designing a policy collection-elimination (PCE) algori
thm and proving a distribution-dependent regret bound that is independent of the
 state-action space. We hope our theoretical results can provide insight towards
 understanding pre-training and generalization in RL.

****
Personalized Federated Learning with Inferred Collaboration Graphs
Rui Ye, Zhenyang Ni, Fangzhao Wu, Siheng Chen, Yanfeng Wang
Personalized federated learning (FL) aims to collaboratively train a personalized model for each client. Previous methods do not adaptively determine who to collaborate at a fine-grained level, making them difficult to handle diverse data heterogeneity levels and those cases where malicious clients exist. To address this issue, our core idea is to learn a collaboration graph, which models the benefits from each pairwise collaboration and allocates appropriate collaboration strengths. Based on this, we propose a novel personalized FL algorithm, pFedGraph, which consists of two key modules: (1) inferring the collaboration graph based on pairwise model similarity and dataset size at server to promote fine-grained collaboration and (2) optimizing local model with the assistance of aggregated model at client to promote personalization. The advantage of pFedGraph is flexibly adaptive to diverse data heterogeneity levels and model poisoning attacks, as the proposed collaboration graph always pushes each client to collaborate more with similar and beneficial clients. Extensive experiments show that pFedGraph consistently outperforms the other 14 baseline methods across various heterogeneity levels and multiple cases where malicious clients exist. Code will be available at https://github.com/MediaBrain-SJTU/pFedGraph.
****
Compositional Exemplars for In-context Learning
Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Tao Yu, Lingpeng Kong
Large pretrained language models (LMs) have shown impressive In-Context Learning (ICL) ability, where the model learns to do an unseen task simply by conditioning on a prompt consisting of input-output examples as demonstration, without any parameter updates. The performance of ICL is highly dominated by the quality of the selected in-context examples. However, previous selection methods are mostly based on simple heuristics, leading to sub-optimal performance. In this work, we systematically formulate in-context example selection as a subset selection problem, and optimize it in an end-to-end fashion. We propose CEIL (Compositional Exemplars for In-context Learning), which is instantiated by Determinantal Point Processes (DPPs) to model the interaction between the given input and in-context examples, and optimized through carefully-designed contrastive learning to obtain preference from LMs. We validate CEIL on 12 classification and generation datasets from 7 distinct NLP tasks, including sentiment analysis, phraphrase detection, natural language inference, commonsense reasoning, open-domain question answering, code generation and semantic parsing. Extensive experiments demonstrate the effectiveness, transferability, compositionality of CEIL, shedding new lights on in-context leaning. Our code is released at https://github.com/HKUNLP/icl-ceil.
****
Corruption-Robust Algorithms with Uncertainty Weighting for Nonlinear Contextual Bandits and Markov Decision Processes
Chenlu Ye, Wei Xiong, Quanquan Gu, Tong Zhang
Despite the significant interest and progress in reinforcement learning (RL) problems with adversarial corruption, current works are either confined to the linear setting or lead to an undesired $\tilde{\mathcal O}(\sqrt{T}\zeta)$ regret bound, where T is the number of rounds and ζ is the total amount of corruption. In this paper, we consider contextual bandits with general function approximation and propose a computationally efficient algorithm to achieve a regret of $\tilde{\mathcal O}(\sqrt{T}+\zeta)$. The proposed algorithm relies on the recently developed uncertainty-weighted least-squares regression from linear contextual bandits (He et al., 2022) and a new weighted estimator of uncertainty for the general function class. In contrast to the existing analysis for the sum of uncertainty that is heavily based on the linear structure, we develop a novel technique to control the sum of weighted uncertainty, thus establishing the final regret bound. We then generalize our algorithm to the episodic MDP and first achieve an additive dependence on the corruption level ζ in the scenario of general function approximation. Notably, our algorithms achieve regret bounds that

either nearly match the lower bound or improve the performance of existing metho
ds for all the corruption levels in both known and unknown ζ cases.
****

GNN&GBDT-Guided Fast Optimizing Framework for Large-scale Integer Programming
Huigen Ye, Hua Xu, Hongyan Wang, Chengming Wang, Yu Jiang
The latest two-stage optimization framework based on graph neural network (GNN)
and large neighborhood search (LNS) is the most popular framework in solving lar
ge-scale integer programs (IPs). However, the framework can not effectively use
the embedding spatial information in GNN and still highly relies on large-scale
solvers in LNS, resulting in the scale of IP being limited by the ability of the
 current solver and performance bottlenecks. To handle these issues, this paper
presents a GNN&GBDT-guided fast optimizing framework for large-scale IPs that on
ly uses a small-scale optimizer to solve large-scale IPs efficiently. Specifical
ly, the proposed framework can be divided into three stages: Multi-task GNN Embe
dding to generate the embedding space, GBDT Prediction to effectively use the em
bedding spatial information, and Neighborhood Optimization to solve large-scale
problems fast using the small-scale optimizer. Extensive experiments show that t
he proposed framework can solve IPs with millions of scales and surpass SCIP and
 Gurobi in the specified wall-clock time using only a small-scale optimizer with
 30% of the problem size. It also shows that the proposed framework can save 99%
 of running time in achieving the same solution quality as SCIP, which verifies
the effectiveness and efficiency of the proposed framework in solving large-scal
e IPs.
****

FedDisco: Federated Learning with Discrepancy-Aware Collaboration
Rui Ye, Mingkai Xu, Jianyu Wang, Chenxin Xu, Siheng Chen, Yanfeng Wang
This work considers the category distribution heterogeneity in federated learnin
g. This issue is due to biased labeling preferences at multiple clients and is a
 typical setting of data heterogeneity. To alleviate this issue, most previous w
orks consider either regularizing local models or fine-tuning the global model,
while they ignore the adjustment of aggregation weights and simply assign weight
s based on the dataset size. However, based on our empirical observations and th
eoretical analysis, we find that the dataset size is not optimal and the discrep
ancy between local and global category distributions could be a beneficial and c
omplementary indicator for determining aggregation weights. We thus propose a no
vel aggregation method, Federated Learning with Discrepancy-Aware Collaboration
(FedDisco), whose aggregation weights not only involve both the dataset size and
 the discrepancy value, but also contribute to a tighter theoretical upper bound
 of the optimization error. FedDisco can promote utility and modularity in a com
munication- and computation-efficient way. Extensive experiments show that our F
edDisco outperforms several state-of-the-art methods and can be easily incorpora
ted with many existing methods to further enhance the performance. Our code will
 be available at https://github.com/MediaBrain-SJTU/FedDisco.
****

Towards Quantum Machine Learning for Constrained Combinatorial Optimization: a Q
uantum QAP Solver
Xinyu Ye, Ge Yan, Junchi Yan
Combinatorial optimization (CO) on the graph is a crucial but challenging resear
ch topic. Recent quantum algorithms provide a new perspective for solving CO pro
blems and have the potential to demonstrate quantum advantage. Quantum Approxima
te Optimization Algorithm (QAOA) is a well-known quantum heuristic for CO constr
ucted by a parametric quantum circuit. However, QAOA is originally designed for
unconstrained problems and the circuit parameters and solutions are jointly solv
ed with time-consuming iterations. In this paper, we propose a novel quantum neu
ral network (QNN) for learning CO problems in a supervised manner to achieve bet
ter and faster results. We focus on the Quadratic Assignment Problem (QAP) with
matching constraints and the node permutation invariance property. To this end,
a quantum neural network called QAP-QNN is devised to translate the QAP into a c
onstrained vertex classification task. Moreover, we study two QAP tasks: Graph M
atching and Traveling Salesman Problem on TorchQauntum simulators, and empirical

ly show the effectiveness of our approach.
****
Temporal Label Smoothing for Early Event Prediction
Hugo Yèche, Alizée Pace, Gunnar Ratsch, Rita Kuznetsova
Models that can predict the occurrence of events ahead of time with low false-al arm rates are critical to the acceptance of decision support systems in the medi cal community. This challenging task is typically treated as a simple binary cla ssification, ignoring temporal dependencies between samples, whereas we propose to exploit this structure. We first introduce a common theoretical framework uni fying dynamic survival analysis and early event prediction. Following an analysi s of objectives from both fields, we propose Temporal Label Smoothing (TLS), a s impler, yet best-performing method that preserves prediction monotonicity over t ime. By focusing the objective on areas with a stronger predictive signal, TLS i mproves performance over all baselines on two large-scale benchmark tasks. Gains are particularly notable along clinically relevant measures, such as event reca ll at low false-alarm rates. TLS reduces the number of missed events by up to a factor of two over previously used approaches in early event prediction.
****
From Temporal to Contemporaneous Iterative Causal Discovery in the Presence of L atent Confounders
Raanan Yehezkel Rohekar, Shami Nisimov, Yaniv Gurwicz, Gal Novik
We present a constraint-based algorithm for learning causal structures from obse rvational time-series data, in the presence of latent confounders. We assume a d iscrete-time, stationary structural vector autoregressive process, with both tem poral and contemporaneous causal relations. One may ask if temporal and contempo raneous relations should be treated differently. The presented algorithm gradual ly refines a causal graph by learning long-term temporal relations before short- term ones, where contemporaneous relations are learned last. This ordering of ca usal relations to be learnt leads to a reduction in the required number of stati stical tests. We validate this reduction empirically and demonstrate that it lea ds to higher accuracy for synthetic data and more plausible causal graphs for re al-world data compared to state-of-the-art algorithms.
****
Doubly Adversarial Federated Bandits
Jialin Yi, Milan Vojnovic
We study a new non-stochastic federated multiarmed bandit problem with multiple agents collaborating via a communication network. The losses of the arms are ass igned by an oblivious adversary that specifies the loss of each arm not only for each time step but also for each agent, which we call doubly adversarial. In th is setting, different agents may choose the same arm in the same time step but o bserve different feedback. The goal of each agent is to find a globally best arm in hindsight that has the lowest cumulative loss averaged over all agents, whic h necessities the communication among agents. We provide regret lower bounds for any federated bandit algorithm under different settings, when agents have acces s to full-information feedback, or the bandit feedback. For the bandit feedback setting, we propose a near-optimal federated bandit algorithm called FEDEXP3. Ou r algorithm gives a positive answer to an open question proposed in (Cesa-Bianch i et al., 2016): FEDEXP3 can guarantee a sub-linear regret without exchanging se quences of selected arm identities or loss sequences among agents. We also provi de numerical evaluations of our algorithm to validate our theoretical results an d demonstrate its effectiveness on synthetic and real-world datasets.
****
Online Prototype Alignment for Few-shot Policy Transfer
Qi Yi, Rui Zhang, Shaohui Peng, Jiaming Guo, Yunkai Gao, Kaizhao Yuan, Ruizhi Ch en, Siming Lan, Xing Hu, Zidong Du, Xishan Zhang, Qi Guo, Yunji Chen
Domain adaptation in RL mainly deals with the changes of observation when transf erring the policy to a new environment. Many traditional approaches of domain ad aptation in RL manage to learn a mapping function between the source and target domain in explicit or implicit ways. However, they typically require access to a bundant data from the target domain. Besides, they often rely on visual clues to

learn the mapping function and may fail when the source domain looks quite different from the target domain. To address these problems, in this paper, we propose a novel framework Online Prototype Alignment (OPA) to learn the mapping function based on the functional similarity of elements and is able to achieve few-shot policy transfer within only several episodes. The key insight of OPA is to introduce an exploration mechanism that can interact with the unseen elements of the target domain in an efficient and purposeful manner, and then connect them with the seen elements in the source domain according to their functionalities (instead of visual clues). Experimental results show that when the target domain looks visually different from the source domain, OPA can achieve better transfer performance even with much fewer samples from the target domain, outperforming prior methods.

****

MonoFlow: Rethinking Divergence GANs via the Perspective of Wasserstein Gradient Flows

Mingxuan Yi, Zhanxing Zhu, Song Liu

The conventional understanding of adversarial training in generative adversarial networks (GANs) is that the discriminator is trained to estimate a divergence, and the generator learns to minimize this divergence. We argue that despite the fact that many variants of GANs were developed following this paradigm, the current theoretical understanding of GANs and their practical algorithms are inconsistent. In this paper, we leverage Wasserstein gradient flows which characterize the evolution of particles in the sample space, to gain theoretical insights and algorithmic inspiration of GANs. We introduce a unified generative modeling framework – MonoFlow: the particle evolution is rescaled via a monotonically increasing mapping of the log density ratio. Under our framework, adversarial training can be viewed as a procedure first obtaining MonoFlow’s vector field via training the discriminator and the generator learns to draw the particle flow defined by the corresponding vector field. We also reveal the fundamental difference between variational divergence minimization and adversarial training. This analysis helps us to identify what types of generator loss functions can lead to the successful training of GANs and suggest that GANs may have more loss designs beyond the literature (e.g., non-saturated loss), as long as they realize MonoFlow. Consistent empirical studies are included to validate the effectiveness of our framework.

****

SE(3) diffusion model with application to protein backbone generation

Jason Yim, Brian L. Trippe, Valentin De Bortoli, Emile Mathieu, Arnaud Doucet, Regina Barzilay, Tommi Jaakkola

The design of novel protein structures remains a challenge in protein engineering for applications across biomedicine and chemistry. In this line of work, a diffusion model over rigid bodies in 3D (referred to as frames) has shown success in generating novel, functional protein backbones that have not been observed in nature. However, there exists no principled methodological framework for diffusion on SE(3), the space of orientation preserving rigid motions in R3, that operates on frames and confers the group invariance. We address these shortcomings by developing theoretical foundations of SE(3) invariant diffusion models on multiple frames followed by a novel framework, FrameDiff, for estimating the SE(3) equivariant score over multiple frames. We apply FrameDiff on monomer backbone generation and find it can generate designable monomers up to 500 amino acids without relying on a pretrained protein structure prediction network that has been integral to previous methods. We find our samples are capable of generalizing beyond any known protein structure.

****

CoCo: A Coupled Contrastive Framework for Unsupervised Domain Adaptive Graph Classification

Nan Yin, Li Shen, Mengzhu Wang, Long Lan, Zeyu Ma, Chong Chen, Xian-Sheng Hua, Xiao Luo

Although graph neural networks (GNNs) have achieved impressive achievements in graph classification, they often need abundant task-specific labels, which could

be extensively costly to acquire. A credible solution is to explore additional l
abeled graphs to enhance unsupervised learning on the target domain. However, ho
w to apply GNNs to domain adaptation remains unsolved owing to the insufficient
exploration of graph topology and the significant domain discrepancy. In this pa
per, we propose Coupled Contrastive Graph Representation Learning (CoCo), which
extracts the topological information from coupled learning branches and reduces
the domain discrepancy with coupled contrastive learning. CoCo contains a graph
convolutional network branch and a hierarchical graph kernel network branch, whi
ch explore graph topology in implicit and explicit manners. Besides, we incorpor
ate coupled branches into a holistic multi-view contrastive learning framework,
which not only incorporates graph representations learned from complementary vie
ws for enhanced understanding, but also encourages the similarity between cross-
domain example pairs with the same semantics for domain alignment. Extensive exp
eriments on popular datasets show that our CoCo outperforms these competing base
lines in different settings generally.
****

Adaptive Estimation of Graphical Models under Total Positivity
Jiaxi Ying, José Vin■■cius De Miranda Cardoso, Daniel P. Palomar
We consider the problem of estimating (diagonally dominant) M-matrices as precis
ion matrices in Gaussian graphical models. Such models have shown interesting pr
operties, e.g., the maximum likelihood estimator exists with as little as two ob
servations in the case of M-matrices, and exists even with one observation in th
e case of diagonally dominant M-matrices. We propose an adaptive multiple-stage
estimation method, which refines the estimate by solving a weighted ℓ_1-reg
ularized problem in each stage. We further design a unified framework based on g
radient projection method to solve the regularized problem, equipped with differ
ent projections to handle the constraints of M-matrices and diagonally dominant
M-matrices. Theoretical analysis of the estimation error is established. The pro
posed method outperforms state-of-the-art methods in estimating precision matric
es and identifying graph edges, as evidenced by synthetic and financial time-ser
ies data sets.
****

Improving Visual Prompt Tuning for Self-supervised Vision Transformers
Seungryong Yoo, Eunji Kim, Dahuin Jung, Jungbeom Lee, Sungroh Yoon
Visual Prompt Tuning (VPT) is an effective tuning method for adapting pretrained
 Vision Transformers (ViTs) to downstream tasks. It leverages extra learnable to
kens, known as prompts, which steer the frozen pretrained ViTs. Although VPT has
 demonstrated its applicability with supervised vision transformers, it often un
derperforms with self-supervised ones. Through empirical observations, we deduce
 that the effectiveness of VPT hinges largely on the ViT blocks with which the p
rompt tokens interact. Specifically, VPT shows improved performance on image cla
ssification tasks for MAE and MoCo v3 when the prompt tokens are inserted into l
ater blocks rather than the first block. These observations suggest that there e
xists an optimal location of blocks for the insertion of prompt tokens. Unfortun
ately, identifying the optimal blocks for prompts within each self-supervised Vi
T for diverse future scenarios is a costly process. To mitigate this problem, we
 propose a simple yet effective method that learns a gate for each ViT block to
adjust its intervention into the prompt tokens. With our method, prompt tokens a
re selectively influenced by blocks that require steering for task adaptation. O
ur method outperforms VPT variants in FGVC and VTAB image classification and ADE
20K semantic segmentation. The code is available at https://github.com/ryongithu
b/GatedPromptTuning.
****

End-to-End Multi-Object Detection with a Regularized Mixture Model
Jaeyoung Yoo, Hojun Lee, Seunghyeon Seo, Inseop Chung, Nojun Kwak
Recent end-to-end multi-object detectors simplify the inference pipeline by remo
ving hand-crafted processes such as non-maximum suppression (NMS). However, duri
ng training, they still heavily rely on heuristics and hand-crafted processes wh
ich deteriorate the reliability of the predicted confidence score. In this paper
, we propose a novel framework to train an end-to-end multi-object detector cons

isting of only two terms: negative log-likelihood (NLL) and a regularization ter
m. In doing so, the multi-object detection problem is treated as density estimat
ion of the ground truth bounding boxes utilizing a regularized mixture density m
odel. The proposed end-to-end multi-object Detection with a Regularized Mixture
Model (D-RMM) is trained by minimizing the NLL with the proposed regularization
term, maximum component maximization (MCM) loss, preventing duplicate prediction
s. Our method reduces the heuristics of the training process and improves the re
liability of the predicted confidence score. Moreover, our D-RMM outperforms the
previous end-to-end detectors on MS COCO dataset. Code is available at https://
github.com/lhj815/D-RMM.
****

EM-Network: Oracle Guided Self-distillation for Sequence Learning
Ji Won Yoon, Sunghwan Ahn, Hyeonseung Lee, Minchan Kim, Seok Min Kim, Nam Soo Ki
m
We introduce EM-Network, a novel self-distillation approach that effectively lev
erages target information for supervised sequence-to-sequence (seq2seq) learning
. In contrast to conventional methods, it is trained with oracle guidance, which
is derived from the target sequence. Since the oracle guidance compactly repres
ents the target-side context that can assist the sequence model in solving the t
ask, the EM-Network achieves a better prediction compared to using only the sour
ce input. To allow the sequence model to inherit the promising capability of the
EM-Network, we propose a new self-distillation strategy, where the original seq
uence model can benefit from the knowledge of the EM-Network in a one-stage mann
er. We conduct comprehensive experiments on two types of seq2seq models: connect
ionist temporal classification (CTC) for speech recognition and attention-based
encoder-decoder (AED) for machine translation. Experimental results demonstrate
that the EM-Network significantly advances the current state-of-the-art approach
es, improving over the best prior work on speech recognition and establishing st
ate-of-the-art performance on WMT’14 and IWSLT’14.
****

Continual Learners are Incremental Model Generalizers
Jaehong Yoon, Sung Ju Hwang, Yue Cao
Motivated by the efficiency and rapid convergence of pre-trained models for solv
ing downstream tasks, this paper extensively studies the impact of Continual Lea
rning (CL) models as pre-trainers. We find that, in both supervised and unsuperv
ised CL, the transfer quality of representations does not show a noticeable degr
adation of fine-tuning performance but rather increases gradually. This is becau
se CL models can learn improved task-general features when easily forgetting tas
k-specific knowledge. Based on this observation, we suggest a new unsupervised C
L framework with masked modeling, which aims to capture fluent task-generic repr
esentation during training. Furthermore, we propose a new fine-tuning scheme, GL
obal Attention Discretization (GLAD), that preserves rich task-generic represent
ation during solving downstream tasks. The model fine-tuned with GLAD achieves c
ompetitive performance and can also be used as a good pre-trained model itself.
We believe this paper breaks the barriers between pre-training and fine-tuning s
teps and leads to a sustainable learning framework in which the continual learne
r incrementally improves model generalization, yielding better transfer to unsee
n tasks.
****

An Investigation into Pre-Training Object-Centric Representations for Reinforcem
ent Learning
Jaesik Yoon, Yi-Fu Wu, Heechul Bae, Sungjin Ahn
Unsupervised object-centric representation (OCR) learning has recently drawn att
ention as a new paradigm of visual representation. This is because of its potent
ial of being an effective pre-training technique for various downstream tasks in
terms of sample efficiency, systematic generalization, and reasoning. Although
image-based reinforcement learning (RL) is one of the most important and thus fr
equently mentioned such downstream tasks, the benefit in RL has surprisingly not
been investigated systematically thus far. Instead, most of the evaluations hav
e focused on rather indirect metrics such as segmentation quality and object pro

perty prediction accuracy. In this paper, we investigate the effectiveness of OC
R pre-training for image-based reinforcement learning via empirical experiments.
 For systematic evaluation, we introduce a simple object-centric visual RL bench
mark and conduct experiments to answer questions such as "Does OCR pre-training
improve performance on object-centric tasks?" and "Can OCR pre-training help wit
h out-of-distribution generalization?". Our results provide empirical evidence f
or valuable insights into the effectiveness of OCR pre-training for RL and the p
otential limitations of its use in certain scenarios. Additionally, this study a
lso examines the critical aspects of incorporating OCR pre-training in RL, inclu
ding performance in a visually complex environment and the appropriate pooling l
ayer to aggregate the object representations.
****

Graph Generative Model for Benchmarking Graph Neural Networks
Minji Yoon, Yue Wu, John Palowitch, Bryan Perozzi, Russ Salakhutdinov
As the field of Graph Neural Networks (GNN) continues to grow, it experiences a
corresponding increase in the need for large, real-world datasets to train and t
est new GNN models on challenging, realistic problems. Unfortunately, such graph
 datasets are often generated from online, highly privacy-restricted ecosystems,
 which makes research and development on these datasets hard, if not impossible.
 This greatly reduces the amount of benchmark graphs available to researchers, c
ausing the field to rely only on a handful of publicly-available datasets. To ad
dress this problem, we introduce a novel graph generative model, Computation Gra
ph Transformer (CGT) that learns and reproduces the distribution of real-world g
raphs in a privacy-controlled way. More specifically, CGT (1) generates effectiv
e benchmark graphs on which GNNs show similar task performance as on the source
graphs, (2) scales to process large-scale graphs, (3) incorporates off-the-shelf
 privacy modules to guarantee end-user privacy of the generated graph. Extensive
 experiments across a vast body of graph generative models show that only our mo
del can successfully generate privacy-controlled, synthetic substitutes of large
-scale real-world graphs that can be effectively used to benchmark GNN models.
****

Analyzing Convergence in Quantum Neural Networks: Deviations from Neural Tangent
 Kernels
Xuchen You, Shouvanik Chakrabarti, Boyang Chen, Xiaodi Wu
A quantum neural network (QNN) is a parameterized mapping efficiently implementa
ble on near-term Noisy Intermediate-Scale Quantum (NISQ) computers. It can be us
ed for supervised learning when combined with classical gradient-based optimizer
s. Despite the existing empirical and theoretical investigations, the convergenc
e of QNN training is not fully understood. Inspired by the success of the neural
 tangent kernels (NTKs) in probing into the dynamics of classical neural network
s, a recent line of works proposes to study over-parameterized QNNs by examining
 a quantum version of tangent kernels. In this work, we study the dynamics of QN
Ns and show that contrary to popular belief it is qualitatively different from t
hat of any kernel regression: due to the unitarity of quantum operations, there
is a non-negligible deviation from the tangent kernel regression derived at the
random initialization. As a result of the deviation, we prove the at-most sublin
ear convergence for QNNs with Pauli measurements, which is beyond the explanator
y power of any kernel regression dynamics. We then present the actual dynamics o
f QNNs in the limit of over-parameterization. The new dynamics capture the chang
e of convergence rate during training and implies that the range of measurements
 is crucial to the fast QNN convergence.
****

Entropy-driven Unsupervised Keypoint Representation Learning in Videos
Ali Younes, Simone Schaub-Meyer, Georgia Chalvatzaki
Extracting informative representations from videos is fundamental for effectivel
y learning various downstream tasks. We present a novel approach for unsupervise
d learning of meaningful representations from videos, leveraging the concept of
image spatial entropy (ISE) that quantifies the per-pixel information in an imag
e. We argue that local entropy of pixel neighborhoods and their temporal evoluti
on create valuable intrinsic supervisory signals for learning prominent features

. Building on this idea, we abstract visual features into a concise representati on of keypoints that act as dynamic information transmitters, and design a deep learning model that learns, purely unsupervised, spatially and temporally consis tent representations directly from video frames. Two original information-theore tic losses, computed from local entropy, guide our model to discover consistent keypoint representations; a loss that maximizes the spatial information covered by the keypoints and a loss that optimizes the keypoints' information transporta tion over time. We compare our keypoint representation to strong baselines for v arious downstream tasks, e.g., learning object dynamics. Our empirical results s how superior performance for our information-driven keypoints that resolve chall enges like attendance to static and dynamic objects or objects abruptly entering and leaving the scene.

****

The Benefits of Model-Based Generalization in Reinforcement Learning

Kenny John Young, Aditya Ramesh, Louis Kirsch, Jürgen Schmidhuber

Model-Based Reinforcement Learning (RL) is widely believed to have the potential to improve sample efficiency by allowing an agent to synthesize large amounts o f imagined experience. Experience Replay (ER) can be considered a simple kind of model, which has proved effective at improving the stability and efficiency of deep RL. In principle, a learned parametric model could improve on ER by general izing from real experience to augment the dataset with additional plausible expe rience. However, given that learned value functions can also generalize, it is n ot immediately obvious why model generalization should be better. Here, we provi de theoretical and empirical insight into when, and how, we can expect data gene rated by a learned model to be useful. First, we provide a simple theorem motiva ting how learning a model as an intermediate step can narrow down the set of pos sible value functions more than learning a value function directly from data usi ng the Bellman equation. Second, we provide an illustrative example showing empi rically how a similar effect occurs in a more concrete setting with neural netwo rk function approximation. Finally, we provide extensive experiments showing the benefit of model-based learning for online RL in environments with combinatoria l complexity, but factored structure that allows a learned model to generalize. In these experiments, we take care to control for other factors in order to isol ate, insofar as possible, the benefit of using experience generated by a learned model relative to ER alone.

****

COLA: Orchestrating Error Coding and Learning for Robust Neural Network Inferenc e Against Hardware Defects

Anlan Yu, Ning Lyu, Jieming Yin, Zhiyuan Yan, Wujie Wen

Error correcting output codes (ECOCs) have been proposed to improve the robustne ss of deep neural networks (DNNs) against hardware defects of DNN hardware accel erators. Unfortunately, existing efforts suffer from drawbacks that would greatl y impact their practicality: 1) robust accuracy (with defects) improvement at th e cost of degraded clean accuracy (without defects); 2) no guarantee on better r obust or clean accuracy using stronger ECOCs. In this paper, we first shed light on the connection between these drawbacks and error correlation, and then propo se a novel comprehensive error decorrelation framework, namely COLA. Specificall y, we propose to reduce inner layer feature error correlation by 1) adopting a s eparated architecture, where the last portions of the paths to all output nodes are separated, and 2) orthogonalizing weights in common DNN layers so that the i ntermediate features are orthogonal with each other. We also propose a regulariz ation technique based on total correlation to mitigate overall error correlation at the outputs. The effectiveness of COLA is first analyzed theoretically, and then evaluated experimentally, e.g. up to 6.7% clean accuracy improvement compar ed with the original DNNs and up to 40% robust accuracy improvement compared to the state-of-the-art ECOC-enhanced DNNs.

****

Delving into Noisy Label Detection with Clean Data

Chenglin Yu, Xinsong Ma, Weiwei Liu

A critical element of learning with noisy labels is noisy label detection. Notab

ly, numerous previous works assume that no source of labels can be clean in a no
isy label detection context. In this work, we relax this assumption and assume t
hat a small subset of the training data is clean, which enables substantial nois
y label detection performance gains. Specifically, we propose a novel framework
that leverages clean data by framing the problem of noisy label detection with c
lean data as a multiple hypothesis testing problem. Moreover, we propose BHN, a
simple yet effective approach for noisy label detection that integrates the Benj
amini-Hochberg (BH) procedure into deep neural networks. BHN achieves $\textit{s
tate-of-the-art}$ performance and outperforms baselines by $\textbf{28.48}$% in
terms of false discovery rate (FDR) and by $\textbf{18.99}$% in terms of F1 on C
IFAR-10. Extensive ablation studies further demonstrate the superiority of BHN.
Our code is available at https://github.com/ChenglinYu/BHN.
****

Bag of Tricks for Training Data Extraction from Language Models
Weichen Yu, Tianyu Pang, Qian Liu, Chao Du, Bingyi Kang, Yan Huang, Min Lin, Shu
icheng Yan
With the advance of language models, privacy protection is receiving more attent
ion. Training data extraction is therefore of great importance, as it can serve
as a potential tool to assess privacy leakage. However, due to the difficulty of
 this task, most of the existing methods are proof-of-concept and still not effe
ctive enough. In this paper, we investigate and benchmark tricks for improving t
raining data extraction using a publicly available dataset. Because most existin
g extraction methods use a pipeline of generating-then-ranking, i.e., generating
 text candidates as potential training data and then ranking them based on speci
fic criteria, our research focuses on the tricks for both text generation (e.g.,
 sampling strategy) and text ranking (e.g., token-level criteria). The experimen
tal results show that several previously overlooked tricks can be crucial to the
 success of training data extraction. Based on the GPT-Neo 1.3B evaluation resul
ts, our proposed tricks outperform the baseline by a large margin in most cases,
 providing a much stronger baseline for future research. The code is available a
t https://github.com/weichen-yu/LM-Extraction.
****

Discover-Then-Rank Unlabeled Support Vectors in the Dual Space for Multi-Class A
ctive Learning
Dayou Yu, Weishi Shi, Qi Yu
We propose to approach active learning (AL) from a novel perspective of discover
ing and then ranking potential support vectors by leveraging the key properties
of the dual space of a sparse kernel max-margin predictor. We theoretically anal
yze the change of a hinge loss in the dual form and provide both the upper and l
ower bounds that are deeply connected to the key geometric properties induced by
 the dual space, which then help us identify various types of important data sam
ples for AL. These bounds inform the design of a novel sampling strategy that le
verages class-wise evidence as a key vehicle, formed through an affine combinati
on of dual variables and kernel evaluation. We construct two distinct types of s
ampling functions, including discovery and ranking. The former focuses on sample
s with low total evidence from all classes, which signifies their potential to s
upport exploration; the latter exploits the current decision boundary to identif
y the most conflicting regions for sampling, aiming to further refine the decisi
on boundary. These two functions, which are complementary to each other, are aut
omatically arranged into a two-phase active sampling process that starts with th
e discovery and then transitions to the ranking of data points to most effective
ly balance exploration and exploitation. Experiments on various real-world data
demonstrate the state-of-the-art AL performance achieved by our model.
****

Long-Term Rhythmic Video Soundtracker
Jiashuo Yu, Yaohui Wang, Xinyuan Chen, Xiao Sun, Yu Qiao
We consider the problem of generating musical soundtracks in sync with rhythmic
visual cues. Most existing works rely on pre-defined music representations, lead
ing to the incompetence of generative flexibility and complexity. Other methods
directly generating video-conditioned waveforms suffer from limited scenarios, s

hort lengths, and unstable generation quality. To this end, we present Long-Term Rhythmic Video Soundtracker (LORIS), a novel framework to synthesize long-term conditional waveforms. Specifically, our framework consists of a latent conditional diffusion probabilistic model to perform waveform synthesis. Furthermore, a series of context-aware conditioning encoders are proposed to take temporal information into consideration for a long-term generation. Notably, we extend our model's applicability from dances to multiple sports scenarios such as floor exercise and figure skating. To perform comprehensive evaluations, we establish a benchmark for rhythmic video soundtracks including the pre-processed dataset, improved evaluation metrics, and robust generative baselines. Extensive experiments show that our model generates long-term soundtracks with state-of-the-art musical quality and rhythmic correspondence. Codes are available at https://github.com/OpenGVLab/LORIS.

****

Adversarial Parameter Attack on Deep Neural Networks

Lijia Yu, Yihan Wang, Xiao-Shan Gao

The parameter perturbation attack is a safety threat to deep learning, where small parameter perturbations are made such that the attacked network gives wrong or desired labels of the adversary to specified inputs. However, such attacks could be detected by the user, because the accuracy of the attacked network will reduce and the network cannot work normally. To make the attack more stealthy, in this paper, the adversarial parameter attack is proposed, in which small perturbations to the parameters of the network are made such that the accuracy of the attacked network does not decrease much, but its robustness against adversarial example attacks becomes much lower. As a consequence, the attacked network performs normally on standard samples, but is much more vulnerable to adversarial attacks. The existence of nearly perfect adversarial parameters under L_∞ norm and L_0 norm is proved under reasonable conditions. Algorithms are given which can be used to produce high quality adversarial parameters for the commonly used networks trained with various robust training methods, in that the robustness of the attacked networks decreases significantly when they are evaluated using various adversarial attack methods.

****

CodeIPPrompt: Intellectual Property Infringement Assessment of Code Language Models

Zhiyuan Yu, Yuhao Wu, Ning Zhang, Chenguang Wang, Yevgeniy Vorobeychik, Chaowei Xiao

Recent advances in large language models (LMs) have facilitated their ability to synthesize programming code. However, they have also raised concerns about intellectual property (IP) rights violations. Despite the significance of this issue, it has been relatively less explored. In this paper, we aim to bridge the gap by presenting CodeIPPrompt, a platform for automatic evaluation of the extent to which code language models may reproduce licensed programs. It comprises two key components: prompts constructed from a licensed code database to elicit LMs to generate IP-violating code, and a measurement tool to evaluate the extent of IP violation of code LMs. We conducted an extensive evaluation of existing open-source code LMs and commercial products and revealed the prevalence of IP violations in all these models. We further identified that the root cause is the substantial proportion of training corpus subject to restrictive licenses, resulting from both intentional inclusion and inconsistent license practice in the real world. To address this issue, we also explored potential mitigation strategies, including fine-tuning and dynamic token filtering. Our study provides a testbed for evaluating the IP violation issues of the existing code generation platforms and stresses the need for a better mitigation strategy.

****

SeedGNN: Graph Neural Network for Supervised Seeded Graph Matching

Liren Yu, Jiaming Xu, Xiaojun Lin

There is a growing interest in designing Graph Neural Networks (GNNs) for seeded graph matching, which aims to match two unlabeled graphs using only topological information and a small set of seed nodes. However, most previous GNNs for this

task use a semi-supervised approach, which requires a large number of seeds and cannot learn knowledge that is transferable to unseen graphs. In contrast, this paper proposes a new supervised approach that can learn from a training set how to match unseen graphs with only a few seeds. Our SeedGNN architecture incorporates several novel designs, inspired by theoretical studies of seeded graph matching: 1) it can learn to compute and use witness-like information from different hops, in a way that can be generalized to graphs of different sizes; 2) it can use easily-matched node-pairs as new seeds to improve the matching in subsequent layers. We evaluate SeedGNN on synthetic and real-world graphs and demonstrate significant performance improvements over both non-learning and learning algorithms in the existing literature. Furthermore, our experiments confirm that the knowledge learned by SeedGNN from training graphs can be generalized to test graphs of different sizes and categories.

****

Efficient and Equivariant Graph Networks for Predicting Quantum Hamiltonian

Haiyang Yu, Zhao Xu, Xiaofeng Qian, Xiaoning Qian, Shuiwang Ji

We consider the prediction of the Hamiltonian matrix, which finds use in quantum chemistry and condensed matter physics. Efficiency and equivariance are two important, but conflicting factors. In this work, we propose a SE(3)-equivariant network, named QHNet, that achieves efficiency and equivariance. Our key advance lies at the innovative design of QHNet architecture, which not only obeys the underlying symmetries, but also enables the reduction of number of tensor products by 92%. In addition, QHNet prevents the exponential growth of channel dimension when more atom types are involved. We perform experiments on MD17 datasets, including four molecular systems. Experimental results show that our QHNet can achieve comparable performance to the state of the art methods at a significantly faster speed. Besides, our QHNet consumes 50% less memory due to its streamlined architecture. Our code is publicly available as part of the AIRS library (https://github.com/divelab/AIRS).

****

On the Global Convergence of Risk-Averse Policy Gradient Methods with Expected Conditional Risk Measures

Xian Yu, Lei Ying

Risk-sensitive reinforcement learning (RL) has become a popular tool to control the risk of uncertain outcomes and ensure reliable performance in various sequential decision-making problems. While policy gradient methods have been developed for risk-sensitive RL, it remains unclear if these methods enjoy the same global convergence guarantees as in the risk-neutral case. In this paper, we consider a class of dynamic time-consistent risk measures, called Expected Conditional Risk Measures (ECRMs), and derive policy gradient updates for ECRM-based objective functions. Under both constrained direct parameterization and unconstrained softmax parameterization, we provide global convergence and iteration complexities of the corresponding risk-averse policy gradient algorithms. We further test risk-averse variants of REINFORCE and actor-critic algorithms to demonstrate the efficacy of our method and the importance of risk control.

****

Actor-Critic Alignment for Offline-to-Online Reinforcement Learning

Zishun Yu, Xinhua Zhang

Deep offline reinforcement learning has recently demonstrated considerable promises in leveraging offline datasets, providing high-quality models that significantly reduce the online interactions required for fine-tuning. However, such a benefit is often diminished due to the marked state-action distribution shift, which causes significant bootstrap error and wipes out the good initial policy. Existing solutions resort to constraining the policy shift or balancing the sample replay based on their online-ness. However, they require online estimation of distribution divergence or density ratio. To avoid such complications, we propose deviating from existing actor-critic approaches that directly transfer the state-action value functions. Instead, we post-process them by aligning with the offline learned policy, so that the Q-values for actions outside the offline policy are also tamed. As a result, the online fine-tuning can be simply performed as

in the standard actor-critic algorithms. We show empirically that the proposed method improves the performance of the fine-tuned robotic agents on various simulated tasks.

****

Master-ASR: Achieving Multilingual Scalability and Low-Resource Adaptation in ASR with Modular Learning

Zhongzhi Yu, Yang Zhang, Kaizhi Qian, Cheng Wan, Yonggan Fu, Yongan Zhang, Yingyan Celine Lin

Despite the impressive performance recently achieved by automatic speech recognition (ASR), we observe two primary challenges that hinder its broader applications: (1) The difficulty of introducing scalability into the model to support more languages with limited training, inference, and storage overhead; (2) The low-resource adaptation ability that enables effective low-resource adaptation while avoiding over fitting and catastrophic forgetting issues. Inspired by recent findings, we hypothesize that we can address the above challenges with modules widely shared across languages. To this end, we propose an ASR framework, dubbed Master-ASR, that, for the first time, simultaneously achieves strong multilingual scalability and low-resource adaptation ability thanks to its modularize-then-assemble strategy. Specifically, Master-ASR learns a small set of generalizable sub-modules and adaptively assembles them for different languages to reduce the multilingual overhead and enable effective knowledge transfer for low-resource adaptation. Extensive experiments and visualizations demonstrate that Master-ASR can effectively discover language similarity and improve multilingual and low-resource ASR performance over state-of-the-art (SOTA) methods, e.g., under multilingual-ASR, our framework achieves a 0.13~2.41 lower character error rate (CER) with 30% smaller inference overhead over SOTA solutions on multilingual ASR and a comparable CER with nearly 100 times fewer trainable parameters over SOTA solutions on low-resource tuning, respectively.

****

Coordinate Descent Methods for Fractional Minimization

Ganzhao Yuan

We consider a class of structured fractional minimization problems, in which the numerator part of the objective is the sum of a differentiable convex function and a convex non-smooth function, while the denominator part is a convex or concave function. This problem is difficult to solve since it is non-convex. By exploiting the structure of the problem, we propose two Coordinate Descent (CD) methods for solving this problem. The proposed methods iteratively solve a one-dimensional subproblem globally, and they are guaranteed to converge to coordinate-wise stationary points. In the case of a convex denominator, under a weak locally bounded non-convexity condition, we prove that the optimality of coordinate-wise stationary point is stronger than that of the standard critical point and directional point. Under additional suitable conditions, CD methods converge Q-linearly to coordinate-wise stationary points. In the case of a concave denominator, we show that any critical point is a global minimum, and CD methods converge to the global minimum with a sublinear convergence rate. We demonstrate the applicability of the proposed methods to some machine learning and signal processing models. Our experiments on real-world data have shown that our method significantly and consistently outperforms existing methods in terms of accuracy.

****

On the Power of Foundation Models

Yang Yuan

With infinitely many high-quality data points, infinite computational power, an infinitely large foundation model with a perfect training algorithm and guaranteed zero generalization error on the pretext task, can the model be used for everything? This question cannot be answered by the existing theory of representation, optimization or generalization, because the issues they mainly investigate are assumed to be nonexistent here. In this paper, we show that category theory provides powerful machinery to answer this question. We have proved three results. The first one limits the power of prompt-based learning, saying that the model can solve a downstream task with prompts if and only if the task is representabl

e. The second one says fine tuning does not have this limit, as a foundation mod el with the minimum required power (up to symmetry) can theoretically solve down stream tasks for the category defined by pretext task, with fine tuning and enou gh resources. Our final result can be seen as a new type of generalization theor em, showing that the foundation model can generate unseen objects from the targe t category (e.g., images) using the structural information from the source categ ory (e.g., texts). Along the way, we provide a categorical framework for supervi sed and self-supervised learning, which might be of independent interest.
****
Automatic Intrinsic Reward Shaping for Exploration in Deep Reinforcement Learnin g
Mingqi Yuan, Bo Li, Xin Jin, Wenjun Zeng
We present AIRS: Automatic Intrinsic Reward Shaping that intelligently and adapt ively provides high-quality intrinsic rewards to enhance exploration in reinforc ement learning (RL). More specifically, AIRS selects shaping function from a pre defined set based on the estimated task return in real-time, providing reliable exploration incentives and alleviating the biased objective problem. Moreover, w e develop an intrinsic reward toolkit to provide efficient and reliable implemen tations of diverse intrinsic reward approaches. We test AIRS on various tasks of MiniGrid, Procgen, and DeepMind Control Suite. Extensive simulation demonstrate s that AIRS can outperform the benchmarking schemes and achieve superior perform ance with simple architecture.
****
Traversing Between Modes in Function Space for Fast Ensembling
Eunggu Yun, Hyungi Lee, Giung Nam, Juho Lee
Deep ensemble is a simple yet powerful way to improve the performance of deep ne ural networks. Under this motivation, recent works on mode connectivity have sho wn that parameters of ensembles are connected by low-loss subspaces, and one can efficiently collect ensemble parameters in those subspaces. While this provides a way to efficiently train ensembles, for inference, multiple forward passes sh ould still be executed using all the ensemble parameters, which often becomes a serious bottleneck for real-world deployment. In this work, we propose a novel f ramework to reduce such costs. Given a low-loss subspace connecting two modes of a neural network, we build an additional neural network that predicts the outpu t of the original neural network evaluated at a certain point in the low-loss su bspace. The additional neural network, which we call a “bridge”, is a lightweigh t network that takes minimal features from the original network and predicts out puts for the low-loss subspace without forward passes through the original netwo rk. We empirically demonstrate that we can indeed train such bridge networks and significantly reduce inference costs with the help of bridge networks.
****
Conformal Prediction with Missing Values
Margaux Zaffran, Aymeric Dieuleveut, Julie Josse, Yaniv Romano
Conformal prediction is a theoretically grounded framework for constructing pred ictive intervals. We study conformal prediction with missing values in the covar iates – a setting that brings new challenges to uncertainty quantification. We f irst show that the marginal coverage guarantee of conformal prediction holds on imputed data for any missingness distribution and almost all imputation function s. However, we emphasize that the average coverage varies depending on the patte rn of missing values: conformal methods tend to construct prediction intervals t hat under-cover the response conditionally to some missing patterns. This motiva tes our novel generalized conformalized quantile regression framework, missing d ata augmentation, which yields prediction intervals that are valid conditionally to the patterns of missing values, despite their exponential number. We then sh ow that a universally consistent quantile regression algorithm trained on the im puted data is Bayes optimal for the pinball risk, thus achieving valid coverage conditionally to any given data point. Moreover, we examine the case of a linear model, which demonstrates the importance of our proposal in overcoming the hete roskedasticity induced by missing values. Using synthetic and data from critical care, we corroborate our theory and report improved performance of our methods.

****
KDEformer: Accelerating Transformers via Kernel Density Estimation
Amir Zandieh, Insu Han, Majid Daliri, Amin Karbasi
Dot-product attention mechanism plays a crucial role in modern deep architectures (e.g., Transformer) for sequence modeling, however, naïve exact computation of this model incurs quadratic time and memory complexities in sequence length, hindering the training of long-sequence models. Critical bottlenecks are due to the computation of partition functions in the denominator of softmax function as well as the multiplication of the softmax matrix with the matrix of values. Our key observation is that the former can be reduced to a variant of the kernel density estimation (KDE) problem, and an efficient KDE solver can be further utilized to accelerate the latter via subsampling-based fast matrix products. Our proposed KDEformer can approximate the attention in sub-quadratic time with provable spectral norm bounds, while all prior results merely provide entry-wise error bounds. Empirically, we verify that KDEformer outperforms other attention approximations in terms of accuracy, memory, and arithmetic operations on various pre-trained models. For instance, on BigGAN image generation we achieve better generative scores than the exact computation with over 4× speedup. For ImageNet classification with T2T-ViT, KDEformer shows over 18× speedup while the accuracy drop is less than 0.5%.
****
Bayesian Estimation of Differential Privacy
Santiago Zanella-Beguelin, Lukas Wutschitz, Shruti Tople, Ahmed Salem, Victor Rühle, Andrew Paverd, Mohammad Naseri, Boris Köpf, Daniel Jones
Algorithms such as Differentially Private SGD enable training machine learning models with formal privacy guarantees. However, because these guarantees hold with respect to unrealistic adversaries, the protection afforded against practical attacks is typically much better. An emerging strand of work empirically estimates the protection afforded by differentially private training as a confidence interval for the privacy budget $\hat{\varepsilon}$ spent with respect to specific threat models. Existing approaches derive confidence intervals for $\hat{\varepsilon}$ from confidence intervals for false positive and false negative rates of membership inference attacks, which requires training an impractically large number of models to get intervals that can be acted upon. We propose a novel, more efficient Bayesian approach that brings privacy estimates within the reach of practitioners. Our approach reduces sample size by computing a posterior for $\hat{\varepsilon}$ (not just a confidence interval) from the joint posterior of the false positive and false negative rates of membership inference attacks. We implement an end-to-end system for privacy estimation that integrates our approach and state-of-the-art membership inference attacks, and evaluate it on text and vision classification tasks. For the same number of samples, we see a reduction in interval width of up to 40% compared to prior work.
****
When is Realizability Sufficient for Off-Policy Reinforcement Learning?
Andrea Zanette
Understanding when reinforcement learning algorithms can make successful off-policy predictions—and when the may fail to do so–remains an open problem. Typically, model-free algorithms for reinforcement learning are analyzed under a condition called Bellman completeness when they operate off-policy with function approximation, unless additional conditions are met. However, Bellman completeness is a requirement that is much stronger than realizability and that is deemed to be too strong to hold in practice. In this work, we relax this structural assumption and analyze the statistical complexity of off-policy reinforcement learning when only realizability holds for the prescribed function class. We establish finite-sample guarantees for off-policy reinforcement learning that are free of the approximation error term known as inherent Bellman error, and that depend on the interplay of three factors. The first two are well known: they are the metric entropy of the function class and the concentrability coefficient that represents the cost of learning off-policy. The third factor is new, and it measures the violation of Bellman completeness, namely the mis-alignment between the chosen fu

nction class and its image through the Bellman operator. Our analysis directly a
pplies to the solution found by temporal difference algorithms when they converg
e.
****

On Distribution Dependent Sub-Logarithmic Query Time of Learned Indexing
Sepanta Zeighami, Cyrus Shahabi
A fundamental problem in data management is to find the elements in an array tha
t match a query. Recently, learned indexes are being extensively used to solve t
his problem, where they learn a model to predict the location of the items in th
e array. They are empirically shown to outperform non-learned methods (e.g., B-t
rees or binary search that answer queries in $O(\log n)$ time) by orders of magn
itude. However, success of learned indexes has not been theoretically justified.
 Only existing attempt shows the same query time of $O(\log n)$, but with a cons
tant factor improvement in space complexity over non-learned methods, under some
 assumptions on data distribution. In this paper, we significantly strengthen th
is result, showing that under mild assumptions on data distribution, and the sam
e space complexity as non-learned methods, learned indexes can answer queries in
 $O(\log\log n)$ expected query time. We also show that allowing for slightly la
rger but still near-linear space overhead, a learned index can achieve $O(1)$ ex
pected query time. Our results theoretically prove learned indexes are orders of
 magnitude faster than non-learned methods, theoretically grounding their empiri
cal success.
****

Sequential Counterfactual Risk Minimization
Houssam Zenati, Eustache Diemert, Matthieu Martin, Julien Mairal, Pierre Gaillar
d
Counterfactual Risk Minimization (CRM) is a framework for dealing with the logge
d bandit feedback problem, where the goal is to improve a logging policy using o
ffline data. In this paper, we explore the case where it is possible to deploy l
earned policies multiple times and acquire new data. We extend the CRM principle
 and its theory to this scenario, which we call "Sequential Counterfactual Risk
Minimization (SCRM)." We introduce a novel counterfactual estimator and identify
 conditions that can improve the performance of CRM in terms of excess risk and
regret rates, by using an analysis similar to restart strategies in accelerated
optimization methods. We also provide an empirical evaluation of our method in b
oth discrete and continuous action settings, and demonstrate the benefits of mul
tiple deployments of CRM.
****

LookupFFN: Making Transformers Compute-lite for CPU inference
Zhanpeng Zeng, Michael Davies, Pranav Pulijala, Karthikeyan Sankaralingam, Vikas
 Singh
While GPU clusters are the de facto choice for training large deep neural networ
k (DNN) models today, several reasons including ease of workflow, security and c
ost have led to efforts investigating whether CPUs may be viable for inference i
n routine use in many sectors of the industry. But the imbalance between the com
pute capabilities of GPUs and CPUs is huge. Motivated by these considerations, w
e study a module which is a workhorse within modern DNN architectures, GEMM base
d Feed Forward Networks (FFNs), and assess the extent to which it can be made co
mpute- (or FLOP-) lite. Specifically, we propose an alternative formulation (we
call it LookupFFN) to GEMM based FFNs inspired by the recent studies of using Lo
cality Sensitive Hashing (LSH) to approximate FFNs. Our formulation recasts most
 essential operations as a memory look-up, leveraging the trade-off between the
two resources on any platform: compute and memory (since CPUs offer it in abunda
nce). For RoBERTa language model pretraining, our formulation achieves similar p
erformance compared to GEMM based FFNs, while dramatically reducing the required
 FLOP. Our development is complemented with a detailed hardware profiling of str
ategies that will maximize efficiency – not just on contemporary hardware but on
 products that will be offered in the near/medium term future. Code is avaiable
at https://github.com/mlpen/LookupFFN.
****

Attribute-Efficient PAC Learning of Low-Degree Polynomial Threshold Functions wi
th Nasty Noise
Shiwei Zeng, Jie Shen
The concept class of low-degree polynomial threshold functions (PTFs) plays a fu
ndamental role in machine learning. In this paper, we study PAC learning of K-
sparse degree-d PTFs on $\mathbb{R}^n$, where any such concept depends only on
 K out of n attributes of the input. Our main contribution is a new algorith
m that runs in time $({nd}/{\epsilon})^{O(d)}$ and under the Gaussian marginal d
istribution, PAC learns the class up to error rate ϵ with $O(\frac{K^{4
d}}{\epsilon^{2d}} \cdot \log^{5d} n)$ samples even when an $\eta \leq O(\epsilo
n^d)$ fraction of them are corrupted by the nasty noise of Bshouty et al. (2002)
, possibly the strongest corruption model. Prior to this work, attribute-efficie
nt robust algorithms are established only for the special case of sparse homogen
eous halfspaces. Our key ingredients are: 1) a structural result that translates
 the attribute sparsity to a sparsity pattern of the Chow vector under the basis
 of Hermite polynomials, and 2) a novel attribute-efficient robust Chow vector e
stimation algorithm which uses exclusively a restricted Frobenius norm to either
 certify a good approximation or to validate a sparsity-induced degree-$2d$ poly
nomial as a filter to detect corrupted samples.
****
Generative Graph Dictionary Learning
Zhichen Zeng, Ruike Zhu, Yinglong Xia, Hanqing Zeng, Hanghang Tong
Dictionary learning, which approximates data samples by a set of shared atoms, i
s a fundamental task in representation learning. However, dictionary learning ov
er graphs, namely graph dictionary learning (GDL), is much more challenging than
 vectorial data as graphs lie in disparate metric spaces. The sparse literature
on GDL formulates the problem from the reconstructive view and often learns line
ar graph embeddings with a high computational cost. In this paper, we propose a
Fused Gromov-Wasserstein (FGW) Mixture Model named FraMe to address the GDL prob
lem from the generative view. Equipped with the graph generation function based
on the radial basis function kernel and FGW distance, FraMe generates nonlinear
embedding spaces, which, as we theoretically proved, provide a good approximatio
n of the original graph spaces. A fast solution is further proposed on top of th
e expectation-maximization algorithm with guaranteed convergence. Extensive expe
riments demonstrate the effectiveness of the obtained node and graph embeddings,
 and our algorithm achieves significant improvements over the state-of-the-art m
ethods.
****
Stabilizing Transformer Training by Preventing Attention Entropy Collapse
Shuangfei Zhai, Tatiana Likhomanenko, Etai Littwin, Dan Busbridge, Jason Ramapur
am, Yizhe Zhang, Jiatao Gu, Joshua M. Susskind
Training stability is of great importance to Transformers. In this work, we inve
stigate the training dynamics of Transformers by examining the evolution of the
attention layers. In particular, we track the attention entropy for each attenti
on head during the course of training, which is a proxy for model sharpness. We
identify a common pattern across different architectures and tasks, where low at
tention entropy is accompanied by high training instability, which can take the
form of oscillating loss or divergence. We denote the pathologically low attenti
on entropy, corresponding to highly concentrated attention scores, as $\textit{e
ntropy collapse}$. As a remedy, we propose σReparam, a simple and efficie
nt solution where we reparametrize all linear layers with spectral normalization
 and an additional learned scalar. We demonstrate that σReparam successfu
lly prevents entropy collapse in the attention layers, promoting more stable tra
ining. Additionally, we prove a tight lower bound of the attention entropy, whic
h decreases exponentially fast with the spectral norm of the attention logits, p
roviding additional motivation for our approach. We conduct experiments with $\s
igma$Reparam on image classification, image self-supervised learning, machine tr
anslation, speech recognition, and language modeling tasks. We show that $\sigma
$Reparam provides stability and robustness with respect to the choice of hyperpa
rameters, going so far as enabling training (a) a Vision Transformer to competit

ive performance without warmup, weight decay, layer normalization or adaptive optimizers; (b) deep architectures in machine translation and (c) speech recognition to competitive performance without warmup and adaptive optimizers. Code is available at https://github.com/apple/ml-sigma-reparam.

****

Offline Learning in Markov Games with General Function Approximation

Yuheng Zhang, Yu Bai, Nan Jiang

We study offline multi-agent reinforcement learning (RL) in Markov games, where the goal is to learn an approximate equilibrium—such as Nash equilibrium and (Coarse) Correlated Equilibrium—from an offline dataset pre-collected from the game. Existing works consider relatively restricted tabular or linear models and handle each equilibria separately. In this work, we provide the first framework for sample-efficient offline learning in Markov games under general function approximation, handling all 3 equilibria in a unified manner. By using Bellman-consistent pessimism, we obtain interval estimation for policies' returns, and use both the upper and the lower bounds to obtain a relaxation on the gap of a candidate policy, which becomes our optimization objective. Our results generalize prior works and provide several additional insights. Importantly, we require a data coverage condition that improves over the recently proposed "unilateral concentrability". Our condition allows selective coverage of deviation policies that optimally trade-off between their greediness (as approximate best responses) and coverage, and we show scenarios where this leads to significantly better guarantees. As a new connection, we also show how our algorithmic framework can subsume seemingly different solution concepts designed for the special case of two-player zero-sum games.

****

Learning useful representations for shifting tasks and distributions

Jianyu Zhang, Leon Bottou

Does the dominant approach to learn representations (as a side effect of optimizing an expected cost for a single training distribution) remain a good approach when we are dealing with multiple distributions? Our thesis is that such scenarios are better served by representations that are richer than those obtained with a single optimization episode. We support this thesis with simple theoretical arguments and with experiments utilizing an apparently näive ensembling technique: concatenating the representations obtained from multiple training episodes using the same data, model, algorithm, and hyper-parameters, but different random seeds. These independently trained networks perform similarly. Yet, in a number of scenarios involving new distributions, the concatenated representation performs substantially better than an equivalently sized network trained with a single training run. This proves that the representations constructed by multiple training episodes are in fact different. Although their concatenation carries little additional information about the training task under the training distribution, it becomes substantially more informative when tasks or distributions change. Meanwhile, a single training episode is unlikely to yield such a redundant representation because the optimization process has no reason to accumulate features that do not incrementally improve the training performance.

****

Nonparametric Iterative Machine Teaching

Chen Zhang, Xiaofeng Cao, Weiyang Liu, Ivor Tsang, James Kwok

In this paper, we consider the problem of Iterative Machine Teaching (IMT), where the teacher provides examples to the learner iteratively such that the learner can achieve fast convergence to a target model. However, existing IMT algorithms are solely based on parameterized families of target models. They mainly focus on convergence in the parameter space, resulting in difficulty when the target models are defined to be functions without dependency on parameters. To address such a limitation, we study a more general task – Nonparametric Iterative Machine Teaching (NIMT), which aims to teach nonparametric target models to learners in an iterative fashion. Unlike parametric IMT that merely operates in the parameter space, we cast NIMT as a functional optimization problem in the function space. To solve it, we propose both random and greedy functional teaching algorithm

s. We obtain the iterative teaching dimension (ITD) of the random teaching algor ithm under proper assumptions, which serves as a uniform upper bound of ITD in N IMT. Further, the greedy teaching algorithm has a significantly lower ITD, which reaches a tighter upper bound of ITD in NIMT. Finally, we verify the correctnes s of our theoretical findings with extensive experiments in nonparametric scenar ios.

****

Matrix Estimation for Individual Fairness

Cindy Zhang, Sarah Huiyi Cen, Devavrat Shah

In recent years, multiple notions of algorithmic fairness have arisen. One such notion is individual fairness (IF), which requires that individuals who are simi lar receive similar treatment. In parallel, matrix estimation (ME) has emerged a s a natural paradigm for handling noisy data with missing values. In this work, we connect the two concepts. We show that pre-processing data using ME can impro ve an algorithm’s IF without sacrificing performance. Specifically, we show that using a popular ME method known as singular value thresholding (SVT) to pre-pro cess the data provides a strong IF guarantee under appropriate conditions. We th en show that, under analogous conditions, SVT pre-processing also yields estimat es that are consistent and approximately minimax optimal. As such, the ME pre-pr ocessing step does not, under the stated conditions, increase the prediction err or of the base algorithm, i.e., does not impose a fairness-performance trade-off . We verify these results on synthetic and real data.

****

Graph Contrastive Backdoor Attacks

Hangfan Zhang, Jinghui Chen, Lu Lin, Jinyuan Jia, Dinghao Wu

Graph Contrastive Learning (GCL) has attracted considerable interest due to its impressive node representation learning capability. Despite the wide application of GCL techniques, little attention has been paid to the security of GCL. In th is paper, we systematically study the vulnerability of GCL in the presence of ma licious backdoor adversaries. In particular, we propose GCBA, the first backdoor attack for graph contrastive learning. GCBA incorporates three attacks: poisoni ng, crafting, and natural backdoor, each targeting one stage of the GCL pipeline . We formulate our attacks as optimization problems and solve them with a novel discrete optimization technique to overcome the discrete nature of graph-structu red data. By extensively evaluating GCBA on multiple datasets and GCL methods, w e show that our attack can achieve high attack success rates while preserving st ealthiness. We further consider potential countermeasures to our attack and conc lude that existing defenses are insufficient to mitigate GCBA. We show that as a complex paradigm involving data and model republishing, GCL is vulnerable to ba ckdoor attacks, and specifically designed defenses are needed to mitigate the ba ckdoor attacks on GCL.

****

Effective Minkowski Dimension of Deep Nonparametric Regression: Function Approxi mation and Statistical Theories

Zixuan Zhang, Minshuo Chen, Mengdi Wang, Wenjing Liao, Tuo Zhao

Existing theories on deep nonparametric regression have shown that when the inpu t data lie on a low-dimensional manifold, deep neural networks can adapt to the intrinsic data structures. In real world applications, such an assumption of dat a lying exactly on a low dimensional manifold is stringent. This paper introduce s a relaxed assumption that the input data are concentrated around a subset of $ \mathbb{R}^d$ denoted by $\mathcal{S}$, and the intrinsic dimension of $\mathcal {S}$ can be characterized by a new complexity notation – effective Minkowski dim ension. We prove that, the sample complexity of deep nonparametric regression on ly depends on the effective Minkowski dimension of $\mathcal{S}$ denoted by p. We further illustrate our theoretical findings by considering nonparametric reg ression with an anisotropic Gaussian random design $N(0,\Sigma)$, where Σ is full rank. When the eigenvalues of Σ have an exponential or polynomia l decay, the effective Minkowski dimension of such an Gaussian random design is $p=\mathcal{O}(\sqrt{\log n})$ or $p=\mathcal{O}(n^\gamma)$, respectively, where n is the sample size and $\gamma\in(0,1)$ is a small constant depending on th

e polynomial decay rate. Our theory shows that, when the manifold assumption doe s not hold, deep neural networks can still adapt to the effective Minkowski dime nsion of the data, and circumvent the curse of the ambient dimensionality for mo derate sample sizes.
****

Tractable Control for Autoregressive Language Generation
Honghua Zhang, Meihua Dang, Nanyun Peng, Guy Van Den Broeck
Despite the success of autoregressive large language models in text generation, it remains a major challenge to generate text that satisfies complex constraints : sampling from the conditional distribution ${\Pr}(\text{text} | \alpha)$ is in tractable for even the simplest lexical constraints α. To overcome this c hallenge, we propose to use tractable probabilistic models (TPMs) to impose lexi cal constraints in autoregressive text generation models, which we refer to as G eLaTo (Generating Language with Tractable Constraints). To demonstrate the effec tiveness of this framework, we use distilled hidden Markov models, where we can efficiently compute ${\Pr}(\text{text} | \alpha)$, to guide autoregressive gener ation from GPT2. GeLaTo achieves state-of-the-art performance on challenging ben chmarks for constrained text generation (e.g., CommonGen), beating various stron g baselines by a large margin. Our work not only opens up new avenues for contro lling large language models but also motivates the development of more expressiv e TPMs.
****

CataBEEM: Integrating Latent Interaction Categories in Node-wise Community Detec tion Models for Network Data
Yuhua Zhang, Walter H. Dempsey
Community detection is a fundamental task in network analysis. Learning underlyi ng network structures has brought deep insights into the understanding of comple x systems. While many methods have focused on clustering nodes into blocks, few accounts for the fact that interactions may exhibit edge-level clustering, which we call categories. Real network data often arise via a series of interactions. Interactions in complex systems can often be clustered into different categorie s and node-level community structures that depend on the category. In this paper , we introduce a category-and-block edge exchangeable model (CataBEEM) to study interaction networks with joint latent interaction-level category and node-level community structures. In particular, the proposed method models the network fro m the interaction process perspective and allows the incorporation of prior know ledge from auxiliary interaction-wise information. We derive an efficient variat ional inference algorithm that can be applied to networks consisting of millions of interactions and provide the theoretical bound of the misspecification rate. We demonstrate the effectiveness of our method in various simulation settings a nd apply the method to TalkLife data, a large-scale online peer-to-peer support network. We show CataBEEM detects more temporally consistent community structure s and has better predictions than other methods.
****

Rethink DARTS Search Space and Renovate a New Benchmark
Jiuling Zhang, Zhiming Ding
DARTS search space (DSS) has become a canonical benchmark for NAS whereas some e merging works pointed out the issue of narrow accuracy range and claimed it woul d hurt the method ranking. We observe some recent studies already suffer from th is issue that overshadows the meaning of scores. In this work, we first propose and orchestrate a suite of improvements to frame a larger and harder DSS, termed LHD, while retaining high efficiency in search. We step forward to renovate a L HD-based new benchmark, taking care of both discernibility and accessibility. Sp ecifically, we re-implement twelve baselines and evaluate them across twelve con ditions by combining two underexpolored influential factors: transductive robust ness and discretization policy, to reasonably construct a benchmark upon multi-c ondition evaluation. Considering that the tabular benchmarks are always insuffic ient to adequately evaluate the methods of neural architecture search (NAS), our work can serve as a crucial basis for the future progress of NAS.
****

Team Belief DAG: Generalizing the Sequence Form to Team Games for Fast Computation of Correlated Team Max-Min Equilibria via Regret Minimization
Brian Hu Zhang, Gabriele Farina, Tuomas Sandholm
A classic result in the theory of extensive-form games asserts that the set of strategies available to any perfect-recall player is strategically equivalent to a low-dimensional convex polytope, called the sequence-form polytope. Online convex optimization tools operating on this polytope are the current state-of-the-art for computing several notions of equilibria in games, and have been crucial in landmark applications of computational game theory. However, when optimizing over the joint strategy space of a team of players, one cannot use the sequence form to obtain a strategically-equivalent convex description of the strategy set of the team. In this paper, we provide new complexity results on the computation of optimal strategies for teams, and propose a new representation, coined team belief DAG (TB-DAG), that describes team strategies as a convex set. The TB-DAG enjoys state-of-the-art parameterized complexity bounds, while at the same time enjoying the advantages of efficient regret minimization techniques. We show that TB-DAG can be exponentially smaller and can be computed exponentially faster than all other known representations, and that the converse is never true. Experimentally, we show that the TB-DAG, when paired with learning techniques, yields state of the art on a wide variety of benchmark team games.
****
A Complete Expressiveness Hierarchy for Subgraph GNNs via Subgraph Weisfeiler-Lehman Tests
Bohang Zhang, Guhao Feng, Yiheng Du, Di He, Liwei Wang
Recently, subgraph GNNs have emerged as an important direction for developing expressive graph neural networks (GNNs). While numerous architectures have been proposed, so far there is still a limited understanding of how various design paradigms differ in terms of expressive power, nor is it clear what design principle achieves maximal expressiveness with minimal architectural complexity. To address these fundamental questions, this paper conducts a systematic study of general node-based subgraph GNNs through the lens of Subgraph Weisfeiler-Lehman Tests (SWL). Our central result is to build a complete hierarchy of SWL with strictly growing expressivity. Concretely, we prove that any node-based subgraph GNN falls into one of the six SWL equivalence classes, among which SSWL achieves the maximal expressive power. We also study how these equivalence classes differ in terms of their practical expressiveness such as encoding graph distance and biconnectivity. In addition, we give a tight expressivity upper bound of all SWL algorithms by establishing a close relation with localized versions of WL and Folklore WL (FWL) tests. Overall, our results provide insights into the power of existing subgraph GNNs, guide the design of new architectures, and point out their limitations by revealing an inherent gap with the 2-FWL test. Finally, experiments demonstrate that SSWL-inspired subgraph GNNs can significantly outperform prior architectures on multiple benchmarks despite great simplicity.
****
Crafting Training Degradation Distribution for the Accuracy-Generalization Trade-off in Real-World Super-Resolution
Ruofan Zhang, Jinjin Gu, Haoyu Chen, Chao Dong, Yulun Zhang, Wenming Yang
Super-resolution (SR) techniques designed for real-world applications commonly encounter two primary challenges: generalization performance and restoration accuracy. We demonstrate that when methods are trained using complex, large-range degradations to enhance generalization, a decline in accuracy is inevitable. However, since the degradation in a certain real-world applications typically exhibits a limited variation range, it becomes feasible to strike a trade-off between generalization performance and testing accuracy within this scope. In this work, we introduce a novel approach to craft training degradation distributions using a small set of reference images. Our strategy is founded upon the binned representation of the degradation space and the Frechet distance between degradation distributions. Our results indicate that the proposed technique significantly improves the performance of test images while preserving generalization capabilities

in real-world applications.
****
Prompting Large Language Model for Machine Translation: A Case Study
Biao Zhang, Barry Haddow, Alexandra Birch
Research on prompting has shown excellent performance with little or even no supervised training across many tasks. However, prompting for machine translation is still under-explored in the literature. We fill this gap by offering a systematic study on prompting strategies for translation, examining various factors for prompt template and demonstration example selection. We further explore the use of monolingual data and the feasibility of cross-lingual, cross-domain, and sentence-to-document transfer learning in prompting. Extensive experiments with GLM-130B (Zeng et al., 2022) as the testbed show that 1) the number and the quality of prompt examples matter, where using suboptimal examples degenerates translation; 2) several features of prompt examples, such as semantic similarity, show significant Spearman correlation with their prompting performance; yet, none of the correlations are strong enough; 3) using pseudo parallel prompt examples constructed from monolingual data via zero-shot prompting could improve translation; and 4) improved performance is achievable by transferring knowledge from prompt examples selected in other settings. We finally provide an analysis on the model outputs and discuss several problems that prompting still suffers from.
****
On the Interplay Between Misspecification and Sub-optimality Gap in Linear Contextual Bandits
Weitong Zhang, Jiafan He, Zhiyuan Fan, Quanquan Gu
We study linear contextual bandits in the misspecified setting, where the expected reward function can be approximated by a linear function class up to a bounded misspecification level $\zeta>0$. We propose an algorithm based on a novel data selection scheme, which only selects the contextual vectors with large uncertainty for online regression. We show that, when the misspecification level ζ is dominated by $\tilde O(\Delta / \sqrt{d})$ with Δ being the minimal sub-optimality gap and d being the dimension of the contextual vectors, our algorithm enjoys the same gap-dependent regret bound $\tilde O ({d^2} /{\Delta})$ as in the well-specified setting up to logarithmic factors. Given this result, we show that the existing SupLinUCB algorithm (Chu et al., 2011) can also achieve a gap-dependent constant regret bound without the knowledge of sub-optimality gap Δ. Together with a lower bound adapted from Lattimore et al. (2020), our result suggests an interplay between the misspecification level and the sub-optimality gap: (1) the linear contextual bandit model is efficiently learnable when $\zeta \leq \tilde O({\Delta} / \sqrt{d})$; and (2) it is not efficiently learnable when $\zeta \geq \tilde \Omega({\Delta} / {\sqrt{d}})$. Experiments on both synthetic and real-world datasets corroborate our theoretical results.
****
When Sparsity Meets Contrastive Models: Less Graph Data Can Bring Better Class-Balanced Representations
Chunhui Zhang, Chao Huang, Yijun Tian, Qianlong Wen, Zhongyu Ouyang, Youhuan Li, Yanfang Ye, Chuxu Zhang
Graph Neural Networks (GNNs) are powerful models for non-Euclidean data, but their training is often accentuated by massive unnecessary computation: on the one hand, training on non-Euclidean data has relatively high computational cost due to its irregular density properties; on the other hand, the class imbalance property often associated with non-Euclidean data cannot be alleviated by the massiveness of the data, thus hindering the generalisation of the models. To address the above issues, theoretically, we start with a hypothesis about the effectiveness of using a subset of training data for GNNs, which is guaranteed by the gradient distance between the subset and the full set. Empirically, we also observe that a subset of the data can provide informative gradients for model optimization and which changes over time dynamically. We name this phenomenon dynamic data sparsity. Additionally, we find that pruned sparse contrastive models may miss valuable information, leading to a large loss value on the informative subset. Motivated by the above findings, we develop a unified data model dynamic sparsity

framework called Data Decantation (DataDec) to address the above challenges. The key idea of DataDec is to identify the informative subset dynamically during the training process by applying sparse graph contrastive learning. The effectiveness of DataDec is comprehensively evaluated on graph benchmark datasets and we also verify its generalizability on image data.

****

Spatial-Temporal Graph Learning with Adversarial Contrastive Adaptation

Qianru Zhang, Chao Huang, Lianghao Xia, Zheng Wang, Siu Ming Yiu, Ruihua Han

Spatial-temporal graph learning has emerged as the state-of-the-art solution for modeling structured spatial-temporal data in learning region representations for various urban sensing tasks (e.g., crime forecasting, traffic flow prediction). However, most existing models are vulnerable to the quality of the generated region graph due to the inartistic graph-structured information aggregation schema. The ubiquitous spatial-temporal data noise and incompleteness in real-life scenarios bring difficulties to generate high-quality region representations. In this paper, we propose a Spatial-Temporal Adversarial Graph contrastive learning model (STAG) to tackle this challenge for adaptive self-supervised graph augmentation. Specifically, we propose a learnable contrastive learning function that enables the automated distillation of important multi-view self-supervised signals for adaptive spatial-temporal graph augmentation. To enhance the representation discrimination ability and robustness, the designed adversarial contrastive learning mechanism empowers STAG to adaptively identify hard samples for better self-supervision. Finally, a cross-view contrastive learning paradigm is introduced to model the inter-dependencies across view-specific region representations and preserve the underlying relation heterogeneity. We verify the superiority of our STAG method in various spatial-temporal prediction tasks on several benchmark datasets.

****

Towards Coherent Image Inpainting Using Denoising Diffusion Implicit Models

Guanhua Zhang, Jiabao Ji, Yang Zhang, Mo Yu, Tommi Jaakkola, Shiyu Chang

Image inpainting refers to the task of generating a complete, natural image based on a partially revealed reference image. Recently, many research interests have been focused on addressing this problem using fixed diffusion models. These approaches typically directly replace the revealed region of the intermediate or final generated images with that of the reference image or its variants. However, since the unrevealed regions are not directly modified to match the context, it results in incoherence between revealed and unrevealed regions. To address the incoherence problem, a small number of methods introduce a rigorous Bayesian framework, but they tend to introduce mismatches between the generated and the reference images due to the approximation errors in computing the posterior distributions. In this paper, we propose CoPaint, which can coherently inpaint the whole image without introducing mismatches. CoPaint also uses the Bayesian framework to jointly modify both revealed and unrevealed regions but approximates the posterior distribution in a way that allows the errors to gradually drop to zero throughout the denoising steps, thus strongly penalizing any mismatches with the reference image. Our experiments verify that CoPaint can outperform the existing diffusion-based methods under both objective and subjective metrics.

****

CAB: Comprehensive Attention Benchmarking on Long Sequence Modeling

Jun Zhang, Shuyang Jiang, Jiangtao Feng, Lin Zheng, Lingpeng Kong

Transformer has achieved remarkable success in language, image, and speech processing. Recently, various efficient attention architectures have been proposed to improve transformer's efficiency while largely preserving its efficacy, especially in modeling long sequences. A widely-used benchmark to test these efficient methods' capability on long-range modeling is Long Range Arena (LRA). However, LRA only focuses on the standard bidirectional (or noncausal) self attention, and completely ignores cross attentions and unidirectional (or causal) attentions, which are equally important to downstream applications. In this paper, we propose Comprehensive Attention Benchmark (CAB) under a fine-grained attention taxonomy with four distinguishable attention patterns, namely, noncausal self, causal s

elf, noncausal cross, and causal cross attentions. CAB collects seven real-world tasks from different research areas to evaluate efficient attentions under the four attention patterns. Among these tasks, CAB validates efficient attentions in eight backbone networks to show their generalization across neural architectures. We conduct exhaustive experiments to benchmark the performances of nine widely-used efficient attention architectures designed with different philosophies on CAB. Extensive experimental results also shed light on the fundamental problems of efficient attentions, such as efficiency length against vanilla attention, performance consistency across attention patterns, the benefit of attention mechanisms, and interpolation/extrapolation on long-context language modeling.
****

Adaptive Barrier Smoothing for First-Order Policy Gradient with Contact Dynamics
Shenao Zhang, Wanxin Jin, Zhaoran Wang

Differentiable physics-based simulators have witnessed remarkable success in robot learning involving contact dynamics, benefiting from their improved accuracy and efficiency in solving the underlying complementarity problem. However, when utilizing the First-Order Policy Gradient (FOPG) method, our theory indicates that the complementarity-based systems suffer from stiffness, leading to an explosion in the gradient variance of FOPG. As a result, optimization becomes challenging due to chaotic and non-smooth loss landscapes. To tackle this issue, we propose a novel approach called Adaptive Barrier Smoothing (ABS), which introduces a class of softened complementarity systems that correspond to barrier-smoothed objectives. With a contact-aware adaptive central-path parameter, ABS reduces the FOPG gradient variance while controlling the gradient bias. We justify the adaptive design by analyzing the roots of the system’s stiffness. Additionally, we establish the convergence of FOPG and show that ABS achieves a reasonable trade-off between the gradient variance and bias by providing their upper bounds. Moreover, we present a variant of FOPG based on complementarity modeling that efficiently fits the contact dynamics by learning the physical parameters. Experimental results on various robotic tasks are provided to support our theory and method.
****

One-Step Estimator for Permuted Sparse Recovery
Hang Zhang, Ping Li

This paper considers the unlabeled sparse recovery under multiple measurements, i.e., ${\mathbf{Y}} = {\mathbf{\Pi}}^{\natural} {\mathbf{X}} {\mathbf{B}}^{\natural} + {\mathbf{W}}$, where ${\mathbf{Y}} \in \mathbb{R}^{n\times m}, {\mathbf{\Pi}}^{\natural}\in \mathbb{R}^{n\times n}, {\mathbf{X}} \in \mathbb{R}^{n\times p}, {\mathbf{B}} ^{\natural}\in \mathbb{R}^{p\times m}, {\mathbf{W}}\in \mathbb{R}^{n\times m}$ represents the observations, missing (or incomplete) correspondence information, sensing matrix, sparse signals, and additive sensing noise, respectively. Different from the previous works on multiple measurements ($m > 1$) which all focus on the sufficient samples regime, namely, $n > p$, we consider a sparse matrix $\mathbf{B}^{\natural}$ and investigate the insufficient samples regime (i.e., $n \ll p$) for the first time. To begin with, we establish the lower bound on the sample number and signal-to-noise ratio ($ {\mathsf{SNR}}$) for the correct permutation recovery. Moreover, we present a simple yet effective estimator. Under mild conditions, we show that our estimator can restore the correct correspondence information with high probability. Numerical experiments are presented to corroborate our theoretical claims.
****

Quantum Lower Bounds for Finding Stationary Points of Nonconvex Functions
Chenyi Zhang, Tongyang Li

Quantum computing is an emerging technology that has been rapidly advancing in the past decades. In this paper, we conduct a systematic study of quantum lower bounds on finding ϵ-approximate stationary points of nonconvex functions, and we consider the following two important settings: 1) having access to p-th order derivatives; or 2) having access to stochastic gradients. The classical query lower bounds are $\Omega\big(\epsilon^{-\frac{1+p}{p}}\big)$ regarding the first setting and $\Omega(\epsilon^{-4})$ regarding the second setting (or $\Omega(\epsilon^{-3})$ if the stochastic gradient function is mean-squared smooth)

. In this paper, we extend all these classical lower bounds to the quantum setting. They match the classical algorithmic results respectively, demonstrating that there is no quantum speedup for finding ϵ-stationary points of nonconvex functions with p-th order derivative inputs or stochastic gradient inputs, whether with or without the mean-squared smoothness assumption. Technically, we prove our quantum lower bounds by showing that the sequential nature of classical hard instances in all these settings also applies to quantum queries, preventing any quantum speedup other than revealing information of the stationary points sequentially.
****

Improving Medical Predictions by Irregular Multimodal Electronic Health Records Modeling
Xinlu Zhang, Shiyang Li, Zhiyu Chen, Xifeng Yan, Linda Ruth Petzold
Health conditions among patients in intensive care units (ICUs) are monitored via electronic health records (EHRs), composed of numerical time series and lengthy clinical note sequences, both taken at $\textit{irregular}$ time intervals. Dealing with such irregularity in every modality, and integrating irregularity into multimodal representations to improve medical predictions, is a challenging problem. Our method first addresses irregularity in each single modality by (1) modeling irregular time series by dynamically incorporating hand-crafted imputation embeddings into learned interpolation embeddings via a gating mechanism, and (2) casting a series of clinical note representations as multivariate irregular time series and tackling irregularity via a time attention mechanism. We further integrate irregularity in multimodal fusion with an interleaved attention mechanism across temporal steps. To the best of our knowledge, this is the first work to thoroughly model irregularity in multimodalities for improving medical predictions. Our proposed methods for two medical prediction tasks consistently outperforms state-of-the-art (SOTA) baselines in each single modality and multimodal fusion scenarios. Specifically, we observe relative improvements of 6.5%, 3.6%, and 4.3% in F1 for time series, clinical notes, and multimodal fusion, respectively. These results demonstrate the effectiveness of our methods and the importance of considering irregularity in multimodal EHRs.
****

FedCR: Personalized Federated Learning Based on Across-Client Common Representation with Conditional Mutual Information Regularization
Hao Zhang, Chenglin Li, Wenrui Dai, Junni Zou, Hongkai Xiong
In personalized federated learning (PFL), multiple clients train customized models to fulfill their personal objectives, which, however, are prone to overfitting to local data due to the heterogeneity and scarcity of local data. To address this, we propose from the information-theoretic perspective a personalized federated learning framework based on the common representation learned across clients, named FedCR. Specifically, we introduce to the local client update a regularizer that aims at minimizing the discrepancy between local and global conditional mutual information (CMI), such that clients are encouraged to learn and exploit the common representation. Upon this, each client learns individually a customized predictor (head), while the extractor (body) remains to be aggregated by the server. Our CMI regularizer leads to a theoretically sound alignment between the local and global stochastic feature distributions in terms of their Kullback-Leibler (KL) divergence. More importantly, by modeling the global joint feature distribution as a product of multiple local feature distributions, clients can efficiently extract diverse information from the global data but without need of the raw data from other clients. We further show that noise injection via feature alignment and ensemble of local predictors in FedCR would help enhance its generalization capability. Experiments on benchmark datasets demonstrate a consistent performance gain and better generalization behavior of FedCR.
****

On the Optimality of Misspecified Kernel Ridge Regression
Haobo Zhang, Yicheng Li, Weihao Lu, Qian Lin
In the misspecified kernel ridge regression problem, researchers usually assume the underground true function $f_{\rho}^{\star} \in [\mathcal{H}]^{s}$, a less-s

mooth interpolation space of a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ for some $s\in (0,1)$. The existing minimax optimal results require $\left\Vert f_{\rho}^{\star} \right \Vert_{L^{\infty}} < \infty$ which implicitly requires $s > \alpha_{0}$ where $\alpha_{0} \in (0,1) $ is the embedding index, a constant depending on $\mathcal{H}$. Whether the KRR is optimal for all $s\in (0,1)$ is an outstanding problem lasting for years. In this paper, we show that KRR is minimax optimal for any $s\in (0,1)$ when the $\mathcal{H}$ is a Sobolev RKHS.
****

Fed-CBS: A Heterogeneity-Aware Client Sampling Mechanism for Federated Learning via Class-Imbalance Reduction
Jianyi Zhang, Ang Li, Minxue Tang, Jingwei Sun, Xiang Chen, Fan Zhang, Changyou Chen, Yiran Chen, Hai Li
Due to the often limited communication bandwidth of edge devices, most existing federated learning (FL) methods randomly select only a subset of devices to participate in training at each communication round. Compared with engaging all the available clients, such a random-selection mechanism could lead to significant performance degradation on non-IID (independent and identically distributed) data. In this paper, we present our key observation that the essential reason resulting in such performance degradation is the class-imbalance of the grouped data from randomly selected clients. Based on this observation, we design an efficient heterogeneity-aware client sampling mechanism, namely, Federated Class-balanced Sampling (Fed-CBS), which can effectively reduce class-imbalance of the grouped dataset from the intentionally selected clients. We first propose a measure of class-imbalance which can be derived in a privacy-preserving way. Based on this measure, we design a computation-efficient client sampling strategy such that the actively selected clients will generate a more class-balanced grouped dataset with theoretical guarantees. Experimental results show that Fed-CBS outperforms the status quo approaches in terms of test accuracy and the rate of convergence while achieving comparable or even better performance than the ideal setting where all the available clients participate in the FL training.
****

Learning Subpocket Prototypes for Generalizable Structure-based Drug Design
Zaixi Zhang, Qi Liu
Generating molecules with high binding affinities to target proteins (a.k.a. structure-based drug design) is a fundamental and challenging task in drug discovery. Recently, deep generative models have achieved remarkable success in generating 3D molecules conditioned on the protein pocket. However, most existing methods consider molecular generation for protein pockets independently while neglecting the underlying connections such as subpocket-level similarities. Subpockets are the local protein environments of ligand fragments and pockets with similar subpockets may bind the same molecular fragment (motif) even though their overall structures are different. Therefore, the trained models can hardly generalize to unseen protein pockets in real-world applications. In this paper, we propose a novel method DrugGPS for generalizable structure-based drug design. With the biochemical priors, we propose to learn subpocket prototypes and construct a global interaction graph to model the interactions between subpocket prototypes and molecular motifs. Moreover, a hierarchical graph transformer encoder and motif-based 3D molecule generation scheme are used to improve the model’s performance. The experimental results show that our model consistently outperforms baselines in generating realistic drug candidates with high affinities in challenging out-of-distribution settings.
****

No One Idles: Efficient Heterogeneous Federated Learning with Parallel Edge and Server Computation
Feilong Zhang, Xianming Liu, Shiyi Lin, Gang Wu, Xiong Zhou, Junjun Jiang, Xiangyang Ji
Federated learning suffers from a latency bottleneck induced by network stragglers, which hampers the training efficiency significantly. In addition, due to the heterogeneous data distribution and security requirements, simple and fast averaging aggregation is not feasible anymore. Instead, complicated aggregation oper

ations, such as knowledge distillation, are required. The time cost for complicated aggregation becomes a new bottleneck that limits the computational efficiency of FL. In this work, we claim that the root cause of training latency actually lies in the aggregation-then-broadcasting workflow of the server. By swapping the computational order of aggregation and broadcasting, we propose a novel and efficient parallel federated learning (PFL) framework that unlocks the edge nodes during global computation and the central server during local computation. This fully asynchronous and parallel pipeline enables handling complex aggregation and network stragglers, allowing flexible device participation as well as achieving scalability in computation. We theoretically prove that synchronous and asynchronous PFL can achieve a similar convergence rate as vanilla FL. Extensive experiments empirically show that our framework brings up to $5.56\times$ speedup compared with traditional FL. Code is available at: https://github.com/Hypervoyager/PFL.

****

The Wisdom of Hindsight Makes Language Models Better Instruction Followers

Tianjun Zhang, Fangchen Liu, Justin Wong, Pieter Abbeel, Joseph E. Gonzalez

Reinforcement learning has seen wide success in finetuning large language models to better align with instructions via human feedback. The so-called algorithm, Reinforcement Learning with Human Feedback (RLHF) demonstrates impressive performance on the GPT series models. However, the underlying reinforcement learning algorithm is complex and requires additional training for reward and value networks. In this paper, we consider an alternative approach: converting feedback to instruction by relabeling the original one and training the model for better alignment in a supervised manner. Such an algorithm doesn’t require any additional parameters except for the original language model and maximally reuses the pretraining pipeline. To achieve this, we formulate instruction alignment problem for language models as a goal-reaching problem in decision making. We propose Hindsight Instruction Relabeling (HIR), a novel algorithm for aligning language models with instructions. The resulting two-stage algorithm shed light to a family of reward-free approaches that utilize the hindsightly relabeled instructions based on feedback. We evaluate the performance of HIR extensively on 12 challenging BigBench reasoning tasks and show that HIR outperforms the baseline algorithms and is comparable to or even surpasses supervised fine-tuning. The implementation of HIR is available at https://github.com/tianjunz/HIR.

****

Detecting Adversarial Data by Probing Multiple Perturbations Using Expected Perturbation Score

Shuhai Zhang, Feng Liu, Jiahao Yang, Yifan Yang, Changsheng Li, Bo Han, Mingkui Tan

Adversarial detection aims to determine whether a given sample is an adversarial one based on the discrepancy between natural and adversarial distributions. Unfortunately, estimating or comparing two data distributions is extremely difficult, especially in high-dimension spaces. Recently, the gradient of log probability density (a.k.a., score) w.r.t. the sample is used as an alternative statistic to compute. However, we find that the score is sensitive in identifying adversarial samples due to insufficient information with one sample only. In this paper, we propose a new statistic called expected perturbation score (EPS), which is essentially the expected score of a sample after various perturbations. Specifically, to obtain adequate information regarding one sample, we perturb it by adding various noises to capture its multi-view observations. We theoretically prove that EPS is a proper statistic to compute the discrepancy between two samples under mild conditions. In practice, we can use a pre-trained diffusion model to estimate EPS for each sample. Last, we pro- pose an EPS-based adversarial detection (EPS- AD) method, in which we develop EPS-based maximum mean discrepancy (MMD) as a metric to measure the discrepancy between the test sample and natural samples. We also prove that the EPS-based MMD between natural and adversarial samples is larger than that among natural samples. Extensive experiments show the superior adversarial detection performance of our EPS-AD.

****

On Enhancing Expressive Power via Compositions of Single Fixed-Size ReLU Network
Shijun Zhang, Jianfeng Lu, Hongkai Zhao
This paper explores the expressive power of deep neural networks through the fra mework of function compositions. We demonstrate that the repeated compositions o f a single fixed-size ReLU network exhibit surprising expressive power, despite the limited expressive capabilities of the individual network itself. Specifical ly, we prove by construction that $\mathcal{L}_2\circ \boldsymbol{g}^{\circ r}\c irc \boldsymbol{\mathcal{L}}_1$ can approximate 1-Lipschitz continuous functio ns on $[0,1]^d$ with an error $\mathcal{O}(r^{-1/d})$, where $\boldsymbol{g}$ is realized by a fixed-size ReLU network, $\boldsymbol{\mathcal{L}}_1$ and $\mathc al{L}_2$ are two affine linear maps matching the dimensions, and $\boldsymbol{g} ^{\circ r}$ denotes the r-times composition of $\boldsymbol{g}$. Furthermore, we extend such a result to generic continuous functions on $[0,1]^d$ with the ap proximation error characterized by the modulus of continuity. Our results reveal that a continuous-depth network generated via a dynamical system has immense ap proximation power even if its dynamics function is time-independent and realized by a fixed-size ReLU network.
****
Bi-directional Masks for Efficient N:M Sparse Training
Yuxin Zhang, Yiting Luo, Mingbao Lin, Yunshan Zhong, Jingjing Xie, Fei Chao, Ron grong Ji
We focus on addressing the dense backward propagation issue for training efficie ncy of N:M fine-grained sparsity that preserves at most N out of M consecutive w eights and achieves practical speedups supported by the N:M sparse tensor core. Therefore, we present a novel method of Bi-directional Masks (Bi-Mask) with its two central innovations in: 1) Separate sparse masks in the two directions of fo rward and backward propagation to obtain training acceleration. It disentangles the forward and backward weight sparsity and overcomes the very dense gradient c omputation. 2) An efficient weight row permutation method to maintain performanc e. It picks up the permutation candidate with the most eligible N:M weight block s in the backward to minimize the gradient gap between traditional unidirectiona l masks and our bi-directional masks. Compared with existing uni-directional sce nario that applies a transposable mask and enables backward acceleration, our Bi -Mask is experimentally demonstrated to be more superior in performance. Also, o ur Bi-Mask performs on par with or even better than methods that fail to achieve backward acceleration. Project of this paper is available at https://github.com /zyxxmu/Bi-Mask.
****
Towards Unbiased Training in Federated Open-world Semi-supervised Learning
Jie Zhang, Xiaosong Ma, Song Guo, Wenchao Xu
Federated Semi-supervised Learning (FedSSL) has emerged as a new paradigm for al lowing distributed clients to collaboratively train a machine learning model ove r scarce labeled data and abundant unlabeled data. However, existing works for F edSSL rely on a closed-world assumption that all local training data and global testing data are from seen classes observed in the labeled dataset. It is crucia l to go one step further: adapting FL models to an open-world setting, where uns een classes exist in the unlabeled data. In this paper, we propose a novel Feder atedopen-world Semi-Supervised Learning (FedoSSL) framework, which can solve the key challenge in distributed and open-world settings, i.e., the biased training process for heterogeneously distributed unseen classes. Specifically, since the advent of a certain unseen class depends on a client basis, the locally unseen classes (exist in multiple clients) are likely to receive differentiated superio r aggregation effects than the globally unseen classes (exist only in one client). We adopt an uncertainty-aware suppressed loss to alleviate the biased trainin g between locally unseen and globally unseen classes. Besides, we enable a calib ration module supplementary to the global aggregation to avoid potential conflic ting knowledge transfer caused by inconsistent data distribution among different clients. The proposed FedoSSL can be easily adapted to state-of-the-art FL meth ods, which is also validated via extensive experiments on benchmarks and real-wo rld datasets (CIFAR-10, CIFAR-100 and CINIC-10).

****
Interactive Object Placement with Reinforcement Learning
Shengping Zhang, Quanling Meng, Qinglin Liu, Liqiang Nie, Bineng Zhong, Xiaopeng Fan, Rongrong Ji
Object placement aims to insert a foreground object into a background image with a suitable location and size to create a natural composition. To predict a diverse distribution of placements, existing methods usually establish a one-to-one mapping from random vectors to the placements. However, these random vectors are not interpretable, which prevents users from interacting with the object placement process. To address this problem, we propose an Interactive Object Placement method with Reinforcement Learning, dubbed IOPRE, to make sequential decisions for producing a reasonable placement given an initial location and size of the foreground. We first design a novel action space to flexibly and stably adjust the location and size of the foreground while preserving its aspect ratio. Then, we propose a multi-factor state representation learning method, which integrates composition image features and sinusoidal positional embeddings of the foreground to make decisions for selecting actions. Finally, we design a hybrid reward function that combines placement assessment and the number of steps to ensure that the agent learns to place objects in the most visually pleasing and semantically appropriate location. Experimental results on the OPA dataset demonstrate that the proposed method achieves state-of-the-art performance in terms of plausibility and diversity.
****
Optimal Shrinkage for Distributed Second-Order Optimization
Fangzhao Zhang, Mert Pilanci
In this work, we address the problem of Hessian inversion bias in distributed second-order optimization algorithms. We introduce a novel shrinkage-based estimator for the resolvent of gram matrices which is asymptotically unbiased, and characterize its non-asymptotic convergence rate in the isotropic case. We apply this estimator to bias correction of Newton steps in distributed second-order optimization algorithms, as well as randomized sketching based methods. We examine the bias present in the naive averaging-based distributed Newton’s method using analytical expressions and contrast it with our proposed biasfree approach. Our approach leads to significant improvements in convergence rate compared to standard baselines and recent proposals, as shown through experiments on both real and synthetic datasets.
****
"Why did the Model Fail?": Attributing Model Performance Changes to Distribution Shifts
Haoran Zhang, Harvineet Singh, Marzyeh Ghassemi, Shalmali Joshi
Machine learning models frequently experience performance drops under distribution shifts. The underlying cause of such shifts may be multiple simultaneous factors such as changes in data quality, differences in specific covariate distributions, or changes in the relationship between label and features. When a model does fail during deployment, attributing performance change to these factors is critical for the model developer to identify the root cause and take mitigating actions. In this work, we introduce the problem of attributing performance differences between environments to distribution shifts in the underlying data generating mechanisms. We formulate the problem as a cooperative game where the players are distributions. We define the value of a set of distributions to be the change in model performance when only this set of distributions has changed between environments, and derive an importance weighting method for computing the value of an arbitrary set of distributions. The contribution of each distribution to the total performance change is then quantified as its Shapley value. We demonstrate the correctness and utility of our method on synthetic, semi-synthetic, and real-world case studies, showing its effectiveness in attributing performance changes to a wide range of distribution shifts.
****
Learning Regions of Interest for Bayesian Optimization with Adaptive Level-Set Estimation

Fengxue Zhang, Jialin Song, James C Bowden, Alexander Ladd, Yisong Yue, Thomas D esautels, Yuxin Chen
We study Bayesian optimization (BO) in high-dimensional and non-stationary scena rios. Existing algorithms for such scenarios typically require extensive hyperpa rameter tuning, which limits their practical effectiveness. We propose a framewo rk, called BALLET, which adaptively filters for a high-confidence region of inte rest (ROI) as a superlevel-set of a nonparametric probabilistic model such as a Gaussian process (GP). Our approach is easy to tune, and is able to focus on loc al region of the optimization space that can be tackled by existing BO methods. The key idea is to use two probabilistic models: a coarse GP to identify the ROI , and a localized GP for optimization within the ROI. We show theoretically that BALLET can efficiently shrink the search space, and can exhibit a tighter regre t bound than standard BO without ROI filtering. We demonstrate empirically the e ffectiveness of BALLET on both synthetic and real-world optimization tasks.
****

A Category-theoretical Meta-analysis of Definitions of Disentanglement
Yivan Zhang, Masashi Sugiyama
Disentangling the factors of variation in data is a fundamental concept in machi ne learning and has been studied in various ways by different researchers, leadi ng to a multitude of definitions. Despite the numerous empirical studies, more t heoretical research is needed to fully understand the defining properties of dis entanglement and how different definitions relate to each other. This paper pres ents a meta-analysis of existing definitions of disentanglement, using category theory as a unifying and rigorous framework. We propose that the concepts of the cartesian and monoidal products should serve as the core of disentanglement. Wi th these core concepts, we show the similarities and crucial differences in deal ing with (i) functions, (ii) equivariant maps, (iii) relations, and (iv) stochas tic maps. Overall, our meta-analysis deepens our understanding of disentanglemen t and its various formulations and can help researchers navigate different defin itions and choose the most appropriate one for their specific context.
****

On the Convergence of SARSA with Linear Function Approximation
Shangtong Zhang, Remi Tachet Des Combes, Romain Laroche
SARSA, a classical on-policy control algorithm for reinforcement learning, is kn own to chatter when combined with linear function approximation: SARSA does not diverge but oscillates in a bounded region. However, little is known about how f ast SARSA converges to that region and how large the region is. In this paper, w e make progress towards this open problem by showing the convergence rate of pro jected SARSA to a bounded region. Importantly, the region is much smaller than t he region that we project into, provided that the the magnitude of the reward is not too large. Existing works regarding the convergence of linear SARSA to a fi xed point all require the Lipschitz constant of SARSA’s policy improvement opera tor to be sufficiently small; our analysis instead applies to arbitrary Lipschit z constants and thus characterizes the behavior of linear SARSA for a new regime .
****

AdaNPC: Exploring Non-Parametric Classifier for Test-Time Adaptation
Yifan Zhang, Xue Wang, Kexin Jin, Kun Yuan, Zhang Zhang, Liang Wang, Rong Jin, T ieniu Tan
Many recent machine learning tasks focus to develop models that can generalize t o unseen distributions. Domain generalization (DG) has become one of the key top ics in various fields. Several literatures show that DG can be arbitrarily hard without exploiting target domain information. To address this issue, test-time a daptive (TTA) methods are proposed. Existing TTA methods require offline target data or extra sophisticated optimization procedures during the inference stage. In this work, we adopt Non-Parametric Classifier to perform the test-time Adapta tion (AdaNPC). In particular, we construct a memory that contains the feature an d label pairs from training domains. During inference, given a test instance, Ad aNPC first recalls k closed samples from the memory to vote for the prediction , and then the test feature and predicted label are added to the memory. In this

way, the sample distribution in the memory can be gradually changed from the tr
aining distribution towards the test distribution with very little extra computa
tion cost. We theoretically justify the rationality behind the proposed method.
Besides, we test our model on extensive numerical experiments. AdaNPC significan
tly outperforms competitive baselines on various DG benchmarks. In particular, w
hen the adaptation target is a series of domains, the adaptation accuracy of Ada
NPC is 50% higher than advanced TTA methods.
****
On the Generalization of Multi-modal Contrastive Learning
Qi Zhang, Yifei Wang, Yisen Wang
Multi-modal contrastive learning (MMCL) has recently garnered considerable inter
est due to its superior performance in visual tasks, achieved by embedding multi
-modal data, such as visual-language pairs. However, there still lack theoretica
l understandings of how MMCL extracts useful visual representation from multi-mo
dal pairs, and particularly, how MMCL outperforms previous approaches like self-
supervised contrastive learning (SSCL). In this paper, by drawing an intrinsic c
onnection between MMCL and asymmetric matrix factorization, we establish the fir
st generalization guarantees of MMCL for visual downstream tasks. Based on this
framework, we further unify MMCL and SSCL by showing that MMCL implicitly perfor
ms SSCL with (pseudo) positive pairs induced by text pairs. Through this unified
perspective, we characterize the advantage of MMCL by showing that text pairs i
nduce more semantically consistent and diverse positive pairs, which, according
to our analysis, provably benefit downstream generalization. Inspired by this fi
nding, we propose several methods to significantly improve the downstream perfor
mance of SSCL on ImageNet by leveraging multi-modal information. Code is availab
le at https://github.com/PKU-ML/CLIP-Help-SimCLR.
****
ConCerNet: A Contrastive Learning Based Framework for Automated Conservation Law
Discovery and Trustworthy Dynamical System Prediction
Wang Zhang, Tsui-Wei Weng, Subhro Das, Alexandre Megretski, Luca Daniel, Lam M.
Nguyen
Deep neural networks (DNN) have shown great capacity of modeling a dynamical sys
tem; nevertheless, they usually do not obey physics constraints such as conserva
tion laws. This paper proposes a new learning framework named $\textbf{ConCerNet
}$ to improve the trustworthiness of the DNN based dynamics modeling to endow th
e invariant properties. $\textbf{ConCerNet}$ consists of two steps: (i) a contra
stive learning method to automatically capture the system invariants (i.e. conse
rvation properties) along the trajectory observations; (ii) a neural projection
layer to guarantee that the learned dynamics models preserve the learned invaria
nts. We theoretically prove the functional relationship between the learned late
nt representation and the unknown system invariant function. Experiments show th
at our method consistently outperforms the baseline neural networks in both coor
dinate error and conservation metrics by a large margin. With neural network bas
ed parameterization and no dependence on prior knowledge, our method can be exte
nded to complex and large-scale dynamics by leveraging an autoencoder.
****
Towards Trustworthy Explanation: On Causal Rationalization
Wenbo Zhang, Tong Wu, Yunlong Wang, Yong Cai, Hengrui Cai
With recent advances in natural language processing, rationalization becomes an
essential self-explaining diagram to disentangle the black box by selecting a su
bset of input texts to account for the major variation in prediction. Yet, exist
ing association-based approaches on rationalization cannot identify true rationa
les when two or more snippets are highly inter-correlated and thus provide a sim
ilar contribution to prediction accuracy, so-called spuriousness. To address thi
s limitation, we novelly leverage two causal desiderata, non-spuriousness and ef
ficiency, into rationalization from the causal inference perspective. We formall
y define a series of probabilities of causation based on a newly proposed struct
ural causal model of rationalization, with its theoretical identification establ
ished as the main component of learning necessary and sufficient rationales. The
superior performance of the proposed causal rationalization is demonstrated on

real-world review and medical datasets with extensive experiments compared to st
ate-of-the-art methods.
****
Demystifying Uneven Vulnerability of Link Stealing Attacks against Graph Neural
Networks
He Zhang, Bang Wu, Shuo Wang, Xiangwen Yang, Minhui Xue, Shirui Pan, Xingliang Y
uan
While graph neural networks (GNNs) dominate the state-of-the-art for exploring g
raphs in real-world applications, they have been shown to be vulnerable to a gro
wing number of privacy attacks. For instance, link stealing is a well-known memb
ership inference attack (MIA) on edges that infers the presence of an edge in a
GNN’s training graph. Recent studies on independent and identically distributed
data (e.g., images) have empirically demonstrated that individuals from differen
t groups suffer from different levels of privacy risks to MIAs, i.e., uneven vul
nerability. However, theoretical evidence of such uneven vulnerability is missin
g. In this paper, we first present theoretical evidence of the uneven vulnerabil
ity of GNNs to link stealing attacks, which lays the foundation for demystifying
 such uneven risks among different groups of edges. We further demonstrate a gro
up-based attack paradigm to expose the practical privacy harm to GNN users deriv
ed from the uneven vulnerability of edges. Finally, we empirically validate the
existence of obvious uneven vulnerability on nine real-world datasets (e.g., abo
ut 25% AUC difference between different groups in the Credit graph). Compared wi
th existing methods, the outperformance of our group-based attack paradigm confi
rms that customising different strategies for different groups results in more e
ffective privacy attacks.
****
Provable Dynamic Fusion for Low-Quality Multimodal Data
Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Z
hou, Xi Peng
The inherent challenge of multimodal fusion is to precisely capture the cross-mo
dal correlation and flexibly conduct cross-modal interaction. To fully release t
he value of each modality and mitigate the influence of low-quality multimodal d
ata, dynamic multimodal fusion emerges as a promising learning paradigm. Despite
 its widespread use, theoretical justifications in this field are still notably
lacking. Can we design a provably robust multimodal fusion method? This paper pr
ovides theoretical understandings to answer this question under a most popular m
ultimodal fusion framework from the generalization perspective. We proceed to re
veal that several uncertainty estimation solutions are naturally available to ac
hieve robust multimodal fusion. Then a novel multimodal fusion framework termed
Quality-aware Multimodal Fusion (QMF) is proposed, which can improve the perform
ance in terms of classification accuracy and model robustness. Extensive experim
ental results on multiple benchmarks can support our findings.
****
ReDi: Efficient Learning-Free Diffusion Inference via Trajectory Retrieval
Kexun Zhang, Xianjun Yang, William Yang Wang, Lei Li
Diffusion models show promising generation capability for a variety of data. Des
pite their high generation quality, the inference for diffusion models is still
time-consuming due to the numerous sampling iterations required. To accelerate t
he inference, we propose ReDi, a simple yet learning-free Retrieval-based Diffus
ion sampling framework. From a precomputed knowledge base, ReDi retrieves a traj
ectory similar to the partially generated trajectory at an early stage of genera
tion, skips a large portion of intermediate steps, and continues sampling from a
 later step in the retrieved trajectory. We theoretically prove that the generat
ion performance of ReDi is guaranteed. Our experiments demonstrate that ReDi imp
roves the model inference efficiency by 2$\times$ speedup. Furthermore, ReDi is
able to generalize well in zero-shot cross-domain image generation such as image
 stylization. The code and demo for ReDi is available at https://github.com/zkx0
6111/ReDiffusion.
****
Nearly Optimal Competitive Ratio for Online Allocation Problems with Two-sided R

esource Constraints and Finite Requests
Qixin Zhang, Wenbing Ye, Zaiyi Chen, Haoyuan Hu, Enhong Chen, Yu Yang
In this paper, we investigate the online allocation problem of maximizing the ov erall revenue subject to both lower and upper bound constraints. Compared to the extensively studied online problems with only resource upper bounds, the two-si ded constraints affect the prospects of resource consumption more severely. As a result, only limited violations of constraints or pessimistic competitive bound s could be guaranteed. To tackle the challenge, we define a measure of feasibili ty ξ^* to evaluate the hardness of this problem, and estimate this measureme nt by an optimization routine with theoretical guarantees. We propose an online algorithm adopting a constructive framework, where we initialize a threshold pri ce vector using the estimation, then dynamically update the price vector and use it for decision-making at each step. It can be shown that the proposed algorith m is $\big(1-O(\frac{\varepsilon}{\xi^*-\varepsilon})\big)$ or $\big(1-O(\frac{\ varepsilon}{\xi^*-\sqrt{\varepsilon}})\big)$ competitive with high probability f or ξ^* known or unknown respectively. To the best of our knowledge, this is the first result establishing a nearly optimal competitive algorithm for solving two-sided constrained online allocation problems with a high probability of fea sibility.
****

Do You Remember? Overcoming Catastrophic Forgetting for Fake Audio Detection
Xiaohui Zhang, Jiangyan Yi, Jianhua Tao, Chenglong Wang, Chu Yuan Zhang
Current fake audio detection algorithms have achieved promising performances on most datasets. However, their performance may be significantly degraded when dea ling with audio of a different dataset. The orthogonal weight modification to ov ercome catastrophic forgetting does not consider the similarity of genuine audio across different datasets. To overcome this limitation, we propose a continual learning algorithm for fake audio detection to overcome catastrophic forgetting, called Regularized Adaptive Weight Modification (RAWM). When fine-tuning a dete ction network, our approach adaptively computes the direction of weight modifica tion according to the ratio of genuine utterances and fake utterances. The adapt ive modification direction ensures the network can effectively detect fake audio on the new dataset while preserving its knowledge of old model, thus mitigating catastrophic forgetting. In addition, genuine audio collected from quite differ ent acoustic conditions may skew their feature distribution, so we introduce a r egularization constraint to force the network to remember the old distribution i n this regard. Our method can easily be generalized to related fields, like spee ch emotion recognition. We also evaluate our approach across multiple datasets a nd obtain a significant performance improvement on cross-dataset experiments.
****

Coder Reviewer Reranking for Code Generation
Tianyi Zhang, Tao Yu, Tatsunori Hashimoto, Mike Lewis, Wen-Tau Yih, Daniel Fried , Sida Wang
Sampling diverse programs from a code language model and reranking with model li kelihood is a popular method for code generation but it is prone to preferring d egenerate solutions. Inspired by collaborative programming, we propose Coder-Rev iewer reranking. We augment Coder language models from past work, which generate programs given language instructions, with Reviewer models, which evaluate the likelihood of the instruction given the generated programs. We perform an extens ive study across six datasets with eight models from three model families. Exper imental results show that Coder-Reviewer reranking leads to consistent and signi ficant improvement (up to 17% absolute accuracy gain) over reranking with the Co der model only. When combined with executability filtering, Coder-Reviewer reran king can often outperform the minimum Bayes risk method. Coder-Reviewer rerankin g is easy to implement by prompting, can generalize to different programming lan guages, and works well with off-the-shelf hyperparameters.
****

DP-Fast MH: Private, Fast, and Accurate Metropolis-Hastings for Large-Scale Baye sian Inference
Wanrong Zhang, Ruqi Zhang

Bayesian inference provides a principled framework for learning from complex dat
a and reasoning under uncertainty. It has been widely applied in machine learnin
g tasks such as medical diagnosis, drug design, and policymaking. In these commo
n applications, data can be highly sensitive. Differential privacy (DP) offers d
ata analysis tools with powerful worst-case privacy guarantees and has been deve
loped as the leading approach in privacy-preserving data analysis. In this paper
, we study Metropolis-Hastings (MH), one of the most fundamental MCMC methods, f
or large-scale Bayesian inference under differential privacy. While most existin
g private MCMC algorithms sacrifice accuracy and efficiency to obtain privacy, w
e provide the first exact and fast DP MH algorithm, using only a minibatch of da
ta in most iterations. We further reveal, for the first time, a three-way trade-
off among privacy, scalability (i.e. the batch size), and efficiency (i.e. the c
onvergence rate), theoretically characterizing how privacy affects the utility a
nd computational cost in Bayesian inference. We empirically demonstrate the effe
ctiveness and efficiency of our algorithm in various experiments.
****

Nearly-tight Bounds for Deep Kernel Learning
Yifan Zhang, Min-Ling Zhang
The generalization analysis of deep kernel learning (DKL) is a crucial and open
problem of kernel methods for deep learning. The implicit nonlinear mapping in D
KL makes existing methods of capacity-based generalization analysis for deep lea
rning invalid. In an attempt to overcome this challenge and make up for the gap
in the generalization theory of DKL, we develop an analysis method based on the
composite relationship of function classes and derive capacity-based bounds with
 mild dependence on the depth, which generalizes learning theory bounds to deep
kernels and serves as theoretical guarantees for the generalization of DKL. In t
his paper, we prove novel and nearly-tight generalization bounds based on the un
iform covering number and the Rademacher chaos complexity for deep (multiple) ke
rnel machines. In addition, for some common classes, we estimate their uniform c
overing numbers and Rademacher chaos complexities by bounding their pseudo-dimen
sions and kernel pseudo-dimensions, respectively. The mild bounds without strong
 assumptions partially explain the good generalization ability of deep learning
combined with kernel methods.
****

OpenFE: Automated Feature Generation with Expert-level Performance
Tianping Zhang, Zheyu Aqa Zhang, Zhiyuan Fan, Haoyan Luo, Fengyuan Liu, Qian Liu
, Wei Cao, Li Jian
The goal of automated feature generation is to liberate machine learning experts
 from the laborious task of manual feature generation, which is crucial for impr
oving the learning performance of tabular data. The major challenge in automated
 feature generation is to efficiently and accurately identify effective features
 from a vast pool of candidate features. In this paper, we present OpenFE, an au
tomated feature generation tool that provides competitive results against machin
e learning experts. OpenFE achieves high efficiency and accuracy with two compon
ents: 1) a novel feature boosting method for accurately evaluating the increment
al performance of candidate features and 2) a two-stage pruning algorithm that p
erforms feature pruning in a coarse-to-fine manner. Extensive experiments on ten
 benchmark datasets show that OpenFE outperforms existing baseline methods by a
large margin. We further evaluate OpenFE in two Kaggle competitions with thousan
ds of data science teams participating. In the two competitions, features genera
ted by OpenFE with a simple baseline model can beat 99.3% and 99.6% data science
 teams respectively. In addition to the empirical results, we provide a theoreti
cal perspective to show that feature generation can be beneficial in a simple ye
t representative setting.
****

Optimal Horizon-Free Reward-Free Exploration for Linear Mixture MDPs
Junkai Zhang, Weitong Zhang, Quanquan Gu
We study reward-free reinforcement learning (RL) with linear function approximat
ion, where the agent works in two phases: (1) in the exploration phase, the agen
t interacts with the environment but cannot access the reward; and (2) in the pl

anning phase, the agent is given a reward function and is expected to find a nea
r-optimal policy based on samples collected in the exploration phase. The sample
 complexities of existing reward-free algorithms have a polynomial dependence on
 the planning horizon, which makes them intractable for long planning horizon RL
 problems. In this paper, we propose a new reward-free algorithm for learning li
near mixture Markov decision processes (MDPs), where the transition probability
 can be parameterized as a linear combination of known feature mappings. At the c
ore of our algorithm is uncertainty-weighted value-targeted regression with expl
oration-driven pseudo-reward and a high-order moment estimator for the aleatoric
 and epistemic uncertainties. When the total reward is bounded by 1, we show t
hat our algorithm only needs to explore $\tilde O\left(d^2\varepsilon^{-2}\righ
t)$ episodes to find an ε-optimal policy, where d is the dimension
 of the feature mapping. The sample complexity of our algorithm only has a polyl
ogarithmic dependence on the planning horizon and therefore is "horizon-free”. I
n addition, we provide an $\Omega\left(d^2\varepsilon^{-2}\right)$ sample comple
xity lower bound, which matches the sample complexity of our algorithm up to log
arithmic factors, suggesting that our algorithm is optimal.
****

Unlocking Slot Attention by Changing Optimal Transport Costs
Yan Zhang, David W. Zhang, Simon Lacoste-Julien, Gertjan J. Burghouts, Cees G. M
. Snoek
Slot attention is a powerful method for object-centric modeling in images and vi
deos. However, its set-equivariance limits its ability to handle videos with a d
ynamic number of objects because it cannot break ties. To overcome this limitati
on, we first establish a connection between slot attention and optimal transport
. Based on this new perspective we propose MESH (Minimize Entropy of Sinkhorn):
a cross-attention module that combines the tiebreaking properties of unregulariz
ed optimal transport with the speed of regularized optimal transport. We evaluat
e slot attention using MESH on multiple object-centric learning benchmarks and f
ind significant improvements over slot attention in every setting.
****

Towards a Persistence Diagram that is Robust to Noise and Varied Densities
Hang Zhang, Kaifeng Zhang, Kai Ming Ting, Ye Zhu
Recent works have identified that existing methods, which construct persistence
 diagrams in Topological Data Analysis (TDA), are not robust to noise and varied
 densities in a point cloud. We analyze the necessary properties of an approach t
hat can address these two issues, and propose a new filter function for TDA base
d on a new data-dependent kernel which possesses these properties. Our empirical
 evaluation reveals that the proposed filter function provides a better means fo
r t-SNE visualization and SVM classification than three existing methods of TDA.
****

Robust Situational Reinforcement Learning in Face of Context Disturbances
Jinpeng Zhang, Yufeng Zheng, Chuheng Zhang, Li Zhao, Lei Song, Yuan Zhou, Jiang
Bian
In many real-world tasks, some parts of state features, called contexts, are ind
ependent of action signals, e.g., customer demand in inventory control, speed of
 lead car in autonomous driving, etc. One of the challenges of reinforcement lea
rning in these applications is that the true context transitions can be easily e
xposed some unknown source of contamination, leading to a shift of context trans
itions between source domains and target domains, which could cause performance
degradation for RL algorithms. However, existing methods on robust RL aim at lea
rning robust policies against the deviations of the entire system dynamics. To t
ackle this problem, this paper proposes the framework of robust situational Mark
ov decision process (RS-MDP) which captures the possible deviations of context t
ransitions explicitly. To scale to large context space, we introduce the softmin
 smoothed robust Bellman operator to learn the robust Q-value approximately, and
 apply our RS-MDP framework to existing RL algorithm SAC to learn the desired ro
bust policies. We conduct experiments on several robot control tasks with dynami
c contexts and inventory control tasks to demonstrate that our algorithm can gen
eralize better and more robust against deviations of context transitions, and ou

tperform existing robust RL algorithms.
****
Patch-level Contrastive Learning via Positional Query for Visual Pre-training
Shaofeng Zhang, Qiang Zhou, Zhibin Wang, Fan Wang, Junchi Yan
Dense contrastive learning (DCL) has been recently explored for learning localized information for dense prediction tasks (e.g., detection and segmentation). It still suffers the difficulty of mining pixels/patches correspondence between two views. A simple way is inputting the same view twice and aligning the pixel/patch representation. However, it would reduce the variance of inputs, and hurts the performance. We propose a plug-in method PQCL (Positional Query for patch-level Contrastive Learning), which allows performing patch-level contrasts between two views with exact patch correspondence. Besides, by using positional queries, PQCL increases the variance of inputs, to enhance training. We apply PQCL to popular transformer-based CL frameworks (DINO and iBOT, and evaluate them on classification, detection and segmentation tasks, where our method obtains stable improvements, especially for dense tasks. It achieves new state-of-the-art in most settings. Code is available at https://github.com/Sherrylone/Query_Contrastive.
****
Men Also Do Laundry: Multi-Attribute Bias Amplification
Dora Zhao, Jerone Andrews, Alice Xiang
The phenomenon of $\textit{bias amplification}$ occurs when models amplify training set biases at test time. Existing metrics measure bias amplification with respect to single annotated attributes (e.g., $\texttt{computer}$). However, large-scale datasets typically consist of instances with multiple attribute annotations (e.g., $\{\texttt{computer}, \texttt{keyboard}\}$). We demonstrate models can learn to exploit correlations with respect to multiple attributes, which are not accounted for by current metrics. Moreover, we show that current metrics can give the erroneous impression that little to no bias amplification has occurred as they aggregate positive and negative bias scores. Further, these metrics lack an ideal value, making them difficult to interpret. To address these shortcomings, we propose a new metric: $\textit{Multi-Attribute Bias Amplification}$. We validate our metric’s utility through a bias amplification analysis on the COCO, imSitu, and CelebA datasets. Finally, we benchmark bias mitigation methods using our proposed metric, suggesting possible avenues for future bias mitigation efforts.
****
Rockmate: an Efficient, Fast, Automatic and Generic Tool for Re-materialization in PyTorch
Xunyi Zhao, Théotime Le Hellard, Lionel Eyraud-Dubois, Julia Gusak, Olivier Beaumont
We propose Rockmate to control the memory requirements when training PyTorch DNN models. Rockmate is an automatic tool that starts from the model code and generates an equivalent model, using a predefined amount of memory for activations, at the cost of a few re-computations. Rockmate automatically detects the structure of computational and data dependencies and rewrites the initial model as a sequence of complex blocks. We show that such a structure is widespread and can be found in many models in the literature (Transformer based models, ResNet, RegNets,...). This structure allows us to solve the problem in a fast and efficient way, using an adaptation of Checkmate (too slow on the whole model but general) at the level of individual blocks and an adaptation of Rotor (fast but limited to sequential models) at the level of the sequence itself. We show through experiments on many models that Rockmate is as fast as Rotor and as efficient as Checkmate, and that it allows in many cases to obtain a significantly lower memory consumption for activations (by a factor of 2 to 5) for a rather negligible overhead (of the order of 10% to 20%). Rockmate is open source and available at https://github.com/topal-team/rockmate.
****
Revisiting Structured Variational Autoencoders
Yixiu Zhao, Scott Linderman
Structured variational autoencoders (SVAEs) combine probabilistic graphical mode

l priors on latent variables, deep neural networks to link latent variables to o bserved data, and structure-exploiting algorithms for approximate posterior infe rence. These models are particularly appealing for sequential data, where the pr ior can capture temporal dependencies. However, despite their conceptual eleganc e, SVAEs have proven difficult to implement, and more general approaches have be en favored in practice. Here, we revisit SVAEs using modern machine learning too ls and demonstrate their advantages over more general alternatives in terms of b oth accuracy and efficiency. First, we develop a modern implementation for hardw are acceleration, parallelization, and automatic differentiation of the message passing algorithms at the core of the SVAE. Second, we show that by exploiting s tructure in the prior, the SVAE learns more accurate models and posterior distri butions, which translate into improved performance on prediction tasks. Third, w e show how the SVAE can naturally handle missing data, and we leverage this abil ity to develop a novel, self-supervised training approach. Altogether, these res ults show that the time is ripe to revisit structured variational autoencoders.
****

On Pitfalls of Test-Time Adaptation
Hao Zhao, Yuejiang Liu, Alexandre Alahi, Tao Lin
Test-Time Adaptation (TTA) has recently gained significant attention as a new pa radigm for tackling distribution shifts. Despite the sheer number of existing me thods, the inconsistent experimental conditions and lack of standardization in p rior literature make it difficult to measure their actual efficacies and progres s. To address this issue, we present a large-scale open-sourced Test-Time Adapta tion Benchmark, dubbed TTAB, which includes nine state-of-the-art algorithms, a diverse array of distribution shifts, and two comprehensive evaluation protocols . Through extensive experiments, we identify three common pitfalls in prior effo rts: (i) choosing appropriate hyper-parameter, especially for model selection, i s exceedingly difficult due to online batch dependency; (ii) the effectiveness o f TTA varies greatly depending on the quality of the model being adapted; (iii) even under optimal algorithmic conditions, existing methods still systematically struggle with certain types of distribution shifts. Our findings suggest that f uture research in the field should be more transparent about their experimental conditions, ensure rigorous evaluations on a broader set of models and shifts, a nd re-examine the assumptions underlying the potential success of TTA for practi cal applications.
****

Addressing Budget Allocation and Revenue Allocation in Data Market Environments Using an Adaptive Sampling Algorithm
Boxin Zhao, Boxiang Lyu, Raul Castro Fernandez, Mladen Kolar
High-quality machine learning models are dependent on access to high-quality tra ining data. When the data are not already available, it is tedious and costly to obtain them. Data markets help with identifying valuable training data: model c onsumers pay to train a model, the market uses that budget to identify data and train the model (the budget allocation problem), and finally the market compensa tes data providers according to their data contribution (revenue allocation prob lem). For example, a bank could pay the data market to access data from other fi nancial institutions to train a fraud detection model. Compensating data contrib utors requires understanding data’s contribution to the model; recent efforts to solve this revenue allocation problem based on the Shapley value are inefficien t to lead to practical data markets. In this paper, we introduce a new algorithm to solve budget allocation and revenue allocation problems simultaneously in li near time. The new algorithm employs an adaptive sampling process that selects d ata from those providers who are contributing the most to the model. Better data means that the algorithm accesses those providers more often, and more frequent accesses corresponds to higher compensation. Furthermore, the algorithm can be deployed in both centralized and federated scenarios, boosting its applicability . We provide theoretical guarantees for the algorithm that show the budget is us ed efficiently and the properties of revenue allocation are similar to Shapley’s . Finally, we conduct an empirical evaluation to show the performance of the alg orithm in practical scenarios and when compared to other baselines. Overall, we

believe that the new algorithm paves the way for the implementation of practical data markets.
****
X-Paste: Revisiting Scalable Copy-Paste for Instance Segmentation using CLIP and StableDiffusion
Hanqing Zhao, Dianmo Sheng, Jianmin Bao, Dongdong Chen, Dong Chen, Fang Wen, Lu Yuan, Ce Liu, Wenbo Zhou, Qi Chu, Weiming Zhang, Nenghai Yu
Copy-Paste is a simple and effective data augmentation strategy for instance segmentation. By randomly pasting object instances onto new background images, it creates new training data for free and significantly boosts the segmentation performance, especially for rare object categories. Although diverse, high-quality object instances used in Copy-Paste result in more performance gain, previous works utilize object instances either from human-annotated instance segmentation datasets or rendered from 3D object models, and both approaches are too expensive to scale up to obtain good diversity. In this paper, we revisit Copy-Paste at scale with the power of newly emerged zero-shot recognition models (e.g., CLIP) and text2image models (e.g., StableDiffusion). We demonstrate for the first time that using a text2image model to generate images or zero-shot recognition model to filter noisily crawled images for different object categories is a feasible way to make Copy-Paste truly scalable. To make such success happen, we design a data acquisition and processing framework, dubbed “X-Paste", upon which a systematic study is conducted. On the LVIS dataset, X-Paste provides impressive improvements over the strong baseline CenterNet2 with Swin-L as the backbone. Specifically, it archives +2.6 box AP and +2.1 mask AP gains on all classes and even more significant gains with +6.8 box AP +6.5 mask AP on long-tail classes.
****
Revisiting Simple Regret: Fast Rates for Returning a Good Arm
Yao Zhao, Connor Stephens, Csaba Szepesvari, Kwang-Sung Jun
Simple regret is a natural and parameter-free performance criterion for pure exploration in multi-armed bandits yet is less popular than the probability of missing the best arm or an ϵ-good arm, perhaps due to lack of easy ways to characterize it. In this paper, we make a significant progress on minimizing simple regret in both data-rich ($T\ge n$) and data-poor regime ($T \le n$) where n is the number of arms and T is the number of samples. At its heart is our improved instance-dependent analysis of the well-known Sequential Halving (SH) algorithm where we bound the probability of returning an arm whose mean reward is not within ϵ from the best (i.e., not ϵ-good) for any choice of $\epsilon>0$, although ϵ is not an input to SH. Our bound not only leads to an optimal worst-case simple regret bound of $\sqrt{n/T}$ up to logarithmic factors but also essentially matches the instance-dependent lower bound for returning an ϵ-good arm reported by Katz-Samuels and Jamieson (2020). For the more challenging data-poor regime, we propose Bracketing SH (BSH) that enjoys the same improvement even without sampling each arm at least once. Our empirical study shows that BSH outperforms existing methods on real-world tasks.
****
Transformed Distribution Matching for Missing Value Imputation
He Zhao, Ke Sun, Amir Dezfouli, Edwin V. Bonilla
We study the problem of imputing missing values in a dataset, which has important applications in many domains. The key to missing value imputation is to capture the data distribution with incomplete samples and impute the missing values accordingly. In this paper, by leveraging the fact that any two batches of data with missing values come from the same data distribution, we propose to impute the missing values of two batches of samples by transforming them into a latent space through deep invertible functions and matching them distributionally. To learn the transformations and impute the missing values simultaneously, a simple and well-motivated algorithm is proposed. Our algorithm has fewer hyperparameters to fine-tune and generates high-quality imputations regardless of how missing values are generated. Extensive experiments over a large number of datasets and competing benchmark algorithms show that our method achieves state-of-the-art performance.

****
Protecting Language Generation Models via Invisible Watermarking
Xuandong Zhao, Yu-Xiang Wang, Lei Li
Language generation models have been an increasingly powerful enabler to many ap plications. Many such models offer free or affordable API access which makes the m potentially vulnerable to model extraction attacks through distillation. To pr otect intellectual property (IP) and make fair use of these models, various tech niques such as lexical watermarking and synonym replacement have been proposed. However, these methods can be nullified by obvious countermeasures such as “syno nym randomization”. To address this issue, we propose GINSW, a novel method to p rotect text generation models from being stolen through distillation. The key id ea of our method is to inject secret signals into the probability vector of the decoding steps for each target token. We can then detect the secret message by p robing a suspect model to tell if it is distilled from the protected one. Experi mental results show that GINSW can effectively identify instances of IP infringe ment with minimal impact on the generation quality of protected APIs. Our method demonstrates an absolute improvement of 19 to 29 points on mean average precisi on (mAP) in detecting suspects compared to previous methods against watermark re moval attacks.
****
Local Optimization Achieves Global Optimality in Multi-Agent Reinforcement Learn ing
Yulai Zhao, Zhuoran Yang, Zhaoran Wang, Jason D. Lee
Policy optimization methods with function approximation are widely used in multi -agent reinforcement learning. However, it remains elusive how to design such al gorithms with statistical guarantees. Leveraging a multi-agent performance diffe rence lemma that characterizes the landscape of multi-agent policy optimization, we find that the localized action value function serves as an ideal descent dir ection for each local policy. Motivated by the observation, we present a multi-a gent PPO algorithm in which the local policy of each agent is updated similarly to vanilla PPO. We prove that with standard regularity conditions on the Markov game and problem-dependent quantities, our algorithm converges to the globally o ptimal policy at a sublinear rate. We extend our algorithm to the off-policy set ting and introduce pessimism to policy evaluation, which aligns with experiments . To our knowledge, this is the first provably convergent multi-agent PPO algori thm in cooperative Markov games.
****
Simplified Temporal Consistency Reinforcement Learning
Yi Zhao, Wenshuai Zhao, Rinu Boney, Juho Kannala, Joni Pajarinen
Reinforcement learning (RL) is able to solve complex sequential decision-making tasks but is currently limited by sample efficiency and required computation. To improve sample efficiency, recent work focuses on model-based RL which interlea ves model learning with planning. Recent methods further utilize policy learning , value estimation, and, self-supervised learning as auxiliary objectives. In th is paper we show that, surprisingly, a simple representation learning approach r elying only on a latent dynamics model trained by latent temporal consistency is sufficient for high-performance RL. This applies when using pure planning with a dynamics model conditioned on the representation, but, also when utilizing the representation as policy and value function features in model-free RL. In exper iments, our approach learns an accurate dynamics model to solve challenging high -dimensional locomotion tasks with online planners while being 4.1$\times$ faste r to train compared to ensemble-based methods. With model-free RL without planni ng, especially on high-dimensional tasks, such as the Deepmind Control Suite Hum anoid and Dog tasks, our approach outperforms model-free methods by a large marg in and matches model-based methods’ sample efficiency while training 2.4$\times$ faster.
****
RLEG: Vision-Language Representation Learning with Diffusion-based Embedding Gen eration
Liming Zhao, Kecheng Zheng, Yun Zheng, Deli Zhao, Jingren Zhou

Vision-language representation learning models (e.g., CLIP) have achieved state-of-the-art performance on various downstream tasks, which usually need large-scale training data to learn discriminative representation. Recent progress on generative diffusion models (e.g., DALL-E 2) has demonstrated that diverse high-quality samples can be synthesized by randomly sampling from generative distribution. By virtue of generative capability in this paper, we propose a novel vision-language Representation Learning method with diffusion-based Embedding Generation (RLEG), which exploits diffusion models to generate feature embedding online for learning effective vision-language representation. Specifically, we first adopt image and text encoders to extract the corresponding embeddings. Secondly, pretrained diffusion-based embedding generators are harnessed to transfer the embedding modality online between vision and language domains. The embeddings generated from the generators are then served as augmented embedding-level samples, which are applied to contrastive learning with the variant of the CLIP framework. Experimental results show that the proposed method could learn effective representation and achieve state-of-the-art performance on various tasks including image classification, image-text retrieval, object detection, semantic segmentation, and text-conditional image generation.

****

Optimal Online Generalized Linear Regression with Stochastic Noise and Its Application to Heteroscedastic Bandits

Heyang Zhao, Dongruo Zhou, Jiafan He, Quanquan Gu

We study the problem of online generalized linear regression in the stochastic setting, where the label is generated from a generalized linear model with possibly unbounded additive noise. We provide a sharp analysis of the classical follow-the-regularized-leader (FTRL) algorithm to cope with the label noise. More specifically, for σ-sub-Gaussian label noise, our analysis provides a regret upper bound of $O(\sigma^2 d \log T) + o(\log T)$, where d is the dimension of the input vector, T is the total number of rounds. We also prove an $\Omega(\sigma^2d\log(T/d))$ lower bound for stochastic online linear regression, which indicates that our upper bound is nearly optimal. In addition, we extend our analysis to a more refined Bernstein noise condition. As an application, we study generalized linear bandits with heterogeneous noise and propose an algorithm based on FTRL to achieve the first variance-aware regret bound.

****

Does Continual Learning Equally Forget All Parameters?

Haiyan Zhao, Tianyi Zhou, Guodong Long, Jing Jiang, Chengqi Zhang

Distribution shift (e.g., task or domain shift) in continual learning (CL) usually results in catastrophic forgetting of previously learned knowledge. Although it can be alleviated by repeatedly replaying buffered data, the every-step replay is time-consuming. In this paper, we study which modules in neural networks are more prone to forgetting by investigating their training dynamics during CL. Our proposed metrics show that only a few modules are more task-specific and sensitive to task change, while others can be shared across tasks as common knowledge. Hence, we attribute forgetting mainly to the former and find that finetuning them only on a small buffer at the end of any CL method can bring non-trivial improvement. Due to the small number of finetuned parameters, such ”Forgetting Prioritized Finetuning (FPF)” is efficient in computation. We further propose a more efficient and simpler method that entirely removes the every-step replay and replaces them by only k-times of FPF periodically triggered during CL. Surprisingly, this ”k-FPF” performs comparably to FPF and outperforms the SOTA CL methods but significantly reduces their computational overhead and cost. In experiments on several benchmarks of class- and domain-incremental CL, FPF consistently improves existing CL methods by a large margin, and k-FPF further excels in efficiency without degrading the accuracy. We also empirically studied the impact of buffer size, epochs per task, and finetuning modules on the cost and accuracy of our methods.

****

Online Learning in Stackelberg Games with an Omniscient Follower

Geng Zhao, Banghua Zhu, Jiantao Jiao, Michael Jordan

We study the problem of online learning in a two-player decentralized cooperative Stackelberg game. In each round, the leader first takes an action, followed by the follower who takes their action after observing the leader’s move. The goal of the leader is to learn to minimize the cumulative regret based on the history of interactions. Differing from the traditional formulation of repeated Stackelberg games, we assume the follower is omniscient, with full knowledge of the true reward, and that they always best-respond to the leader’s actions. We analyze the sample complexity of regret minimization in this repeated Stackelberg game. We show that depending on the reward structure, the existence of the omniscient follower may change the sample complexity drastically, from constant to exponential, even for linear cooperative Stackelberg games. This poses unique challenges for the learning process of the leader and the subsequent regret analysis.
****

Structure-informed Language Models Are Protein Designers
Zaixiang Zheng, Yifan Deng, Dongyu Xue, Yi Zhou, Fei Ye, Quanquan Gu
This paper demonstrates that language models are strong structure-based protein designers. We present LM-Design, a generic approach to reprogramming sequence-based protein language models (pLMs), that have learned massive sequential evolutionary knowledge from the universe of natural protein sequences, to acquire an immediate capability to design preferable protein sequences for given folds. We conduct a structural surgery on pLMs, where a lightweight structural adapter is implanted into pLMs and endows it with structural awareness. During inference, iterative refinement is performed to effectively optimize the generated protein sequences. Experiments show that LM-Design improves the state-of-the-art results by a large margin, leading to 4% to 12% accuracy gains in sequence recovery (e.g., 55.65%/56.63% on CATH 4.2/4.3 single-chain benchmarks, and $>$60% when designing protein complexes). We provide extensive and in-depth analyses, which verify that LM-Design can (1) indeed leverage both structural and sequential knowledge to accurately handle structurally non-deterministic regions, (2) benefit from scaling data and model size, and (3) generalize to other proteins (e.g., antibodies and de novo proteins).
****

Semi-Supervised Offline Reinforcement Learning with Action-Free Trajectories
Qinqing Zheng, Mikael Henaff, Brandon Amos, Aditya Grover
Natural agents can effectively learn from multiple data sources that differ in size, quality, and types of measurements. We study this heterogeneity in the context of offline reinforcement learning (RL) by introducing a new, practically motivated semi-supervised setting. Here, an agent has access to two sets of trajectories: labelled trajectories containing state, action and reward triplets at every timestep, along with unlabelled trajectories that contain only state and reward information. For this setting, we develop and study a simple meta-algorithmic pipeline that learns an inverse dynamics model on the labelled data to obtain proxy-labels for the unlabelled data, followed by the use of any offline RL algorithm on the true and proxy-labelled trajectories. Empirically, we find this simple pipeline to be highly successful — on several D4RL benchmarks (Fu et al., 2020), certain offline RL algorithms can match the performance of variants trained on a fully labelled dataset even when we label only 10% of trajectories which are highly suboptimal. To strengthen our understanding, we perform a large-scale controlled empirical study investigating the interplay of data-centric properties of the labelled and unlabelled datasets, with algorithmic design choices (e.g., choice of inverse dynamics, offline RL algorithm) to identify general trends and best practices for training RL agents on semi-supervised offline datasets.
****

Improved Techniques for Maximum Likelihood Estimation for Diffusion ODEs
Kaiwen Zheng, Cheng Lu, Jianfei Chen, Jun Zhu
Diffusion models have exhibited excellent performance in various domains. The probability flow ordinary differential equation (ODE) of diffusion models (i.e., diffusion ODEs) is a particular case of continuous normalizing flows (CNFs), which enables deterministic inference and exact likelihood evaluation. However, the likelihood estimation results by diffusion ODEs are still far from those of the

state-of-the-art likelihood-based generative models. In this work, we propose several improved techniques for maximum likelihood estimation for diffusion ODEs, including both training and evaluation perspectives. For training, we propose velocity parameterization and explore variance reduction techniques for faster convergence. We also derive an error-bounded high-order flow matching objective for finetuning, which improves the ODE likelihood and smooths its trajectory. For evaluation, we propose a novel training-free truncated-normal dequantization to fill the training-evaluation gap commonly existing in diffusion ODEs. Building upon these techniques, we achieve state-of-the-art likelihood estimation results on image datasets (2.56 on CIFAR-10, 3.43/3.69 on ImageNet-32) without variational dequantization or data augmentation.

****

Fast Sampling of Diffusion Models via Operator Learning

Hongkai Zheng, Weili Nie, Arash Vahdat, Kamyar Azizzadenesheli, Anima Anandkumar

Diffusion models have found widespread adoption in various areas. However, their sampling process is slow because it requires hundreds to thousands of network evaluations to emulate a continuous process defined by differential equations. In this work, we use neural operators, an efficient method to solve the probability flow differential equations, to accelerate the sampling process of diffusion models. Compared to other fast sampling methods that have a sequential nature, we are the first to propose a parallel decoding method that generates images with only one model forward pass. We propose diffusion model sampling with neural operator (DSNO) that maps the initial condition, i.e., Gaussian distribution, to the continuous-time solution trajectory of the reverse diffusion process. To model the temporal correlations along the trajectory, we introduce temporal convolution layers that are parameterized in the Fourier space into the given diffusion model backbone. We show our method achieves state-of-the-art FID of 3.78 for CIFAR-10 and 7.83 for ImageNet-64 in the one-model-evaluation setting.

****

Outline, Then Details: Syntactically Guided Coarse-To-Fine Code Generation

Wenqing Zheng, S P Sharan, Ajay Kumar Jaiswal, Kevin Wang, Yihan Xi, Dejia Xu, Zhangyang Wang

For a complicated algorithm, its implementation by a human programmer usually starts with outlining a rough control flow followed by iterative enrichments, eventually yielding carefully generated syntactic structures and variables in a hierarchy. However, state-of-the-art large language models generate codes in a single pass, without intermediate warm-ups to reflect the structured thought process of "outline-then-detail". Inspired by the recent success of chain-of-thought prompting, we propose ChainCoder, a program synthesis language model that generates Python code progressively, i.e. from coarse to fine in multiple passes. We first decompose source code into layout frame components and accessory components via abstract syntax tree parsing to construct a hierarchical representation. We then reform our prediction target into a multi-pass objective, each pass generates a subsequence, which is concatenated in the hierarchy. Finally, a tailored transformer architecture is leveraged to jointly encode the natural language descriptions and syntactically aligned I/O data samples. Extensive evaluations show that ChainCoder outperforms state-of-the-arts, demonstrating that our progressive generation eases the reasoning procedure and guides the language model to generate higher-quality solutions. Our codes are available at: https://github.com/VITA-Group/ChainCoder.

****

Revisiting Discriminative vs. Generative Classifiers: Theory and Implications

Chenyu Zheng, Guoqiang Wu, Fan Bao, Yue Cao, Chongxuan Li, Jun Zhu

A large-scale deep model pre-trained on massive labeled or unlabeled data transfers well to downstream tasks. Linear evaluation freezes parameters in the pre-trained model and trains a linear classifier separately, which is efficient and attractive for transfer. However, little work has investigated the classifier in linear evaluation except for the default logistic regression. Inspired by the statistical efficiency of naive Bayes, the paper revisits the classical topic on discriminative vs. generative classifiers. Theoretically, the paper considers the

surrogate loss instead of the zero-one loss in analyses and generalizes the classical results from binary cases to multiclass ones. We show that, under mild assumptions, multiclass naive Bayes requires $O(\log n)$ samples to approach its asymptotic error while the corresponding multiclass logistic regression requires $O(n)$ samples, where n is the feature dimension. To establish it, we present a multiclass $\mathcal{H}$-consistency bound framework and an explicit bound for logistic loss, which are of independent interests. Simulation results on a mixture of Gaussian validate our theoretical findings. Experiments on various pre-trained deep vision models show that naive Bayes consistently converges faster as the number of data increases. Besides, naive Bayes shows promise in few-shot cases and we observe the "two regimes" phenomenon in pre-trained supervised models. Our code is available at https://github.com/ML-GSAI/Revisiting-Dis-vs-Gen-Classifiers.

****

Evidential Interactive Learning for Medical Image Captioning

Ervine Zheng, Qi Yu

Medical image captioning alleviates the burden of physicians and possibly reduces medical errors by automatically generating text descriptions to describe image contents and convey findings. It is more challenging than conventional image captioning due to the complexity of medical images and the difficulty of aligning image regions with medical terms. In this paper, we propose an evidential interactive learning framework that leverages evidence-based uncertainty estimation and interactive machine learning to improve image captioning with limited labeled data. The interactive learning process involves three stages: keyword prediction, caption generation, and model retraining. First, the model predicts a list of keywords with evidence-based uncertainty and selects the most informative keywords to seek user feedback. Second, user-approved keywords are used as model input to guide the model to generate satisfactory captions. Third, the model is updated based on user-approved keywords and captions, where evidence-based uncertainty is used to allocate different weights to different data instances. Experiments on two medical image datasets illustrate that the proposed framework can effectively learn from human feedback and improve the model’s performance in the future.

****

Finding the Missing-half: Graph Complementary Learning for Homophily-prone and Heterophily-prone Graphs

Yizhen Zheng, He Zhang, Vincent Lee, Yu Zheng, Xiao Wang, Shirui Pan

Real-world graphs generally have only one kind of tendency in their connections. These connections are either homophilic-prone or heterophily-prone. While graphs with homophily-prone edges tend to connect nodes with the same class (i.e., intra-class nodes), heterophily-prone edges tend to build relationships between nodes with different classes (i.e., inter-class nodes). Existing GNNs only take the original graph as input during training. The problem with this approach is that it forgets to take into consideration the ”missing-half” structural information, that is, heterophily-prone topology for homophily-prone graphs and homophily-prone topology for heterophily-prone graphs. In our paper, we introduce Graph cOmplementAry Learning, namely GOAL, which consists of two components: graph complementation and complemented graph convolution. The first component finds the missing-half structural information for a given graph to complement it. The complemented graph has two sets of graphs including both homophily- and heterophily-prone topology. In the latter component, to handle complemented graphs, we design a new graph convolution from the perspective of optimisation. The experiment results show that GOAL consistently outperforms all baselines in eight real-world datasets.

****

Multi-agent Online Scheduling: MMS Allocations for Indivisible Items

Shengwei Zhou, Rufan Bai, Xiaowei Wu

We consider the problem of fairly allocating a sequence of indivisible items that arrive online in an arbitrary order to a group of n agents with additive normalized valuation functions, we consider the allocation of goods and chores sepa

rately and propose algorithms for approximating maximin share (MMS) allocations for both settings. When agents have identical valuation functions the problem coincides with the semi-online machine covering problem (when items are goods) and load balancing problem (when items are chores), for both of which optimal competitive ratios have been achieved. In this paper we consider the case when agents have general additive valuation functions. For the allocation of goods we show that no competitive algorithm exists even when there are only three agents and propose an optimal 0.5-competitive algorithm for the case of two agents. For the allocation of chores we propose a $(2-1/n)$-competitive algorithm for $n\geq 3$ agents and a $\sqrt{2}\approx 1.414$-competitive algorithm for two agents. Additionally, we show that no algorithm can do better than $15/11\approx 1.364$-competitive for two agents.

****

Eliminating Adversarial Noise via Information Discard and Robust Representation Restoration

Dawei Zhou, Yukun Chen, Nannan Wang, Decheng Liu, Xinbo Gao, Tongliang Liu

Deep neural networks (DNNs) are vulnerable to adversarial noise. Denoising model-based defense is a major protection strategy. However, denoising models may fail and induce negative effects in fully white-box scenarios. In this work, we start from the latent inherent properties of adversarial samples to break the limitations. Unlike solely learning a mapping from adversarial samples to natural samples, we aim to achieve denoising by destroying the spatial characteristics of adversarial noise and preserving the robust features of natural information. Motivated by this, we propose a defense based on information discard and robust representation restoration. Our method utilize complementary masks to disrupt adversarial noise and guided denoising models to restore robust-predictive representations from masked samples. Experimental results show that our method has competitive performance against white-box attacks and effectively reverses the negative effect of denoising models.

****

Brainformers: Trading Simplicity for Efficiency

Yanqi Zhou, Nan Du, Yanping Huang, Daiyi Peng, Chang Lan, Da Huang, Siamak Shakeri, David So, Andrew M. Dai, Yifeng Lu, Zhifeng Chen, Quoc V Le, Claire Cui, James Laudon, Jeff Dean

Transformers are central to recent successes in natural language processing and computer vision. Transformers have a mostly uniform backbone where layers alternate between feed-forward and self-attention in order to build a deep network. Here we investigate this design choice and find that more complex blocks that have different permutations of layer primitives can be more efficient. Using this insight, we develop a complex block, named Brainformer, that consists of a diverse sets of layers such as sparsely gated feed-forward layers, dense feed-forward layers, attention layers, and various forms of layer normalization and activation functions. Brainformer consistently outperforms the state-of-the-art dense and sparse Transformers, in terms of both quality and efficiency. A Brainformer model with 8 billion activated parameters per token demonstrates 2x faster training convergence and 5x faster step time compared to its GLaM counterpart. In downstream task evaluation, Brainformer also demonstrates a 3% higher SuperGLUE score with fine-tuning compared to GLaM with a similar number of activated parameters. Finally, Brainformer largely outperforms a Primer dense model derived with NAS with similar computation per token on fewshot evaluations.

****

Implicit Regularization Leads to Benign Overfitting for Sparse Linear Regression

Mo Zhou, Rong Ge

In deep learning, often the training process finds an interpolator (a solution with 0 training loss), but the test loss is still low. This phenomenon, known as benign overfitting, is a major mystery that received a lot of recent attention. One common mechanism for benign overfitting is implicit regularization, where the training process leads to additional properties for the interpolator, often characterized by minimizing certain norms. However, even for a simple sparse linear regression problem $y = \beta^{\ast\top} x +\xi$ with sparse $\beta^{\ast}$, n

either minimum ℓ_1 or ℓ_2 norm interpolator gives the optimal test los s. In this work, we give a different parametrization of the model which leads to a new implicit regularization effect that combines the benefit of ℓ_1 and ℓ_2 interpolators. We show that training our new model via gradient descent leads to an interpolator with near-optimal test loss. Our result is based on ca reful analysis of the training dynamics and provides another example of implicit regularization effect that goes beyond norm minimization.
****
ODS: Test-Time Adaptation in the Presence of Open-World Data Shift
Zhi Zhou, Lan-Zhe Guo, Lin-Han Jia, Dingchu Zhang, Yu-Feng Li
Test-time adaptation (TTA) adapts a source model to the distribution shift in te sting data without using any source data. There have been plenty of algorithms c oncentrated on covariate shift in the last decade, i.e., $\mathcal{D}_t(X)$, the distribution of the test data is different from the source data. Nonetheless, i n real application scenarios, it is necessary to consider the influence of label distribution shift, i.e., both $\mathcal{D}_t(X)$ and $\mathcal{D}_t(Y)$ are sh ifted, which has not been sufficiently explored yet. To remedy this, we study a new problem setup, namely, TTA with Open-world Data Shift (AODS). The goal of AO DS is simultaneously adapting a model to covariate and label distribution shifts in the test phase. In this paper, we first analyze the relationship between cla ssification error and distribution shifts. Motivated by this, we hence propose a new framework, namely ODS, which decouples the mixed distribution shift and the n addresses covariate and label distribution shifts accordingly. We conduct expe riments on multiple benchmarks with different types of shifts, and the results d emonstrate the superior performance of our method against the state of the arts. Moreover, ODS is suitable for many TTA algorithms.
****
Fourmer: An Efficient Global Modeling Paradigm for Image Restoration
Man Zhou, Jie Huang, Chun-Le Guo, Chongyi Li
Global modeling-based image restoration frameworks have become popular. However, they often require a high memory footprint and do not consider task-specific de gradation. Our work presents an alternative approach to global modeling that is more efficient for image restoration. The key insights which motivate our study are two-fold: 1) Fourier transform is capable of disentangling image degradation and content component to a certain extent, serving as the image degradation pri or, and 2) Fourier domain innately embraces global properties, where each pixel in the Fourier space is involved with all spatial pixels. While adhering to the “spatial interaction + channel evolution” rule of previous studies, we customize the core designs with Fourier spatial interaction modeling and Fourier channel evolution. Our paradigm, Fourmer, achieves competitive performance on common ima ge restoration tasks such as image de-raining, image enhancement, image dehazing , and guided image super-resolution, while requiring fewer computational resourc es. The code for Fourmer will be made publicly available.
****
Controlled Text Generation with Natural Language Instructions
Wangchunshu Zhou, Yuchen Eleanor Jiang, Ethan Wilcox, Ryan Cotterell, Mrinmaya S achan
Large language models can be prompted to pro- duce fluent output for a wide rang e of tasks without being specifically trained to do so. Nevertheless, it is noto riously difficult to control their generation in such a way that it satisfies us er-specified constraints. In this paper, we present InstructCTG, a simple contro lled text generation framework that incorporates different constraints by verbal izing them as natural language instructions. We annotate natural texts through a combination of off-the-shelf NLP tools and simple heuristics with the linguisti c and extra-linguistic constraints they satisfy. Then, we verbalize the constrai nts into natural language instructions to form weakly supervised training data, i.e., we prepend the natural language verbalizations of the constraints in front of their corresponding natural language sentences. Next, we fine-tune a pre-tra ined language model on the augmented corpus. Compared to existing methods, Instr uctCTG is more flexible in terms of the types of constraints it allows the pract

itioner to use. It also does not require any modification of the decoding proced ure. Finally, InstructCTG allows the model to adapt to new constraints without r e-training through the use of in-context learning.
****

NNSplitter: An Active Defense Solution for DNN Model via Automated Weight Obfusc ation
Tong Zhou, Yukui Luo, Shaolei Ren, Xiaolin Xu
As a type of valuable intellectual property (IP), deep neural network (DNN) mode ls have been protected by techniques like watermarking. However, such passive mo del protection cannot fully prevent model abuse. In this work, we propose an act ive model IP protection scheme, namely NNSplitter, which actively protects the m odel by splitting it into two parts: the obfuscated model that performs poorly d ue to weight obfuscation, and the model secrets consisting of the indexes and or iginal values of the obfuscated weights, which can only be accessed by authorize d users with the support of the trusted execution environment. Experimental resu lts demonstrate the effectiveness of NNSplitter, e.g., by only modifying 275 out of over 11 million (i.e., 0.002%) weights, the accuracy of the obfuscated ResNe t-18 model on CIFAR-10 can drop to 10%. Moreover, NNSplitter is stealthy and res ilient against norm clipping and fine-tuning attacks, making it an appealing sol ution for DNN model protection. The code is available at: https://github.com/Ton gzhou0101/NNSplitter.
****

Deep Latent State Space Models for Time-Series Generation
Linqi Zhou, Michael Poli, Winnie Xu, Stefano Massaroli, Stefano Ermon
Methods based on ordinary differential equations (ODEs) are widely used to build generative models of time-series. In addition to high computational overhead du e to explicitly computing hidden states recurrence, existing ODE-based models fa ll short in learning sequence data with sharp transitions - common in many real- world systems - due to numerical challenges during optimization. In this work, w e propose LS4, a generative model for sequences with latent variables evolving a ccording to a state space ODE to increase modeling capacity. Inspired by recent deep state space models (S4), we achieve speedups by leveraging a convolutional representation of LS4 which bypasses the explicit evaluation of hidden states. W e show that LS4 significantly outperforms previous continuous-time generative mo dels in terms of marginal distribution, classification, and prediction scores on real-world datasets in the Monash Forecasting Repository, and is capable of mod eling highly stochastic data with sharp temporal transitions. LS4 sets state-of- the-art for continuous-time latent generative models, with significant improveme nt of mean squared error and tighter variational lower bounds on irregularly-sam pled datasets, while also being x100 faster than other baselines on long sequenc es.
****

SlotGAT: Slot-based Message Passing for Heterogeneous Graphs
Ziang Zhou, Jieming Shi, Renchi Yang, Yuanhang Zou, Qing Li
Heterogeneous graphs are ubiquitous to model complex data. There are urgent need s on powerful heterogeneous graph neural networks to effectively support importa nt applications. We identify a potential semantic mixing issue in existing messa ge passing processes, where the representations of the neighbors of a node v are forced to be transformed to the feature space of v for aggregation, though the neighbors are in different types. That is, the semantics in different node types are entangled together into node v's representation. To address the issue, we p ropose SlotGAT with separate message passing processes in slots, one for each no de type, to maintain the representations in their own node-type feature spaces. Moreover, in a slot-based message passing layer, we design an attention mechanis m for effective slot-wise message aggregation. Further, we develop a slot attent ion technique after the last layer of SlotGAT, to learn the importance of differ ent slots in downstream tasks. Our analysis indicates that the slots in SlotGAT can preserve different semantics in various feature spaces. The superiority of S lotGAT is evaluated against 13 baselines on 6 datasets for node classification a nd link prediction. Our code is at https://github.com/scottjiao/SlotGAT_ICML23/.

****
Fast Online Node Labeling for Very Large Graphs
Baojian Zhou, Yifan Sun, Reza Babanezhad Harikandeh
This paper studies the online node classification problem under a transductive l earning setting. Current methods either invert a graph kernel matrix with $\math cal{O}(n^3)$ runtime and $\mathcal{O}(n^2)$ space complexity or sample a large v olume of random spanning trees, thus are difficult to scale to large graphs. In this work, we propose an improvement based on the online relaxation technique in troduced by a series of works (Rakhlin et al., 2012; Rakhlin & Sridharan, 2015; 2017). We first prove an effective regret $\mathcal{O}(\sqrt{n^{1+\gamma}})$ whe n suitable parameterized graph kernels are chosen, then propose an approximate a lgorithm FastONL enjoying $\mathcal{O}(k\sqrt{n^{1+\gamma}})$ regret based on th is relaxation. The key of FastONL is a generalized local push method that effect ively approximates inverse matrix columns and applies to a series of popular ker nels. Furthermore, the per-prediction cost is $\mathcal{O}(\operatorname{vol}{\m athcal{S}}\log 1/\epsilon)$ locally dependent on the graph with linear memory co st. Experiments show that our scalable method enjoys a better tradeoff between l ocal and global consistency.
****
Horizon-Free and Variance-Dependent Reinforcement Learning for Latent Markov Dec ision Processes
Runlong Zhou, Ruosong Wang, Simon Shaolei Du
We study regret minimization for reinforcement learning (RL) in Latent Markov De cision Processes (LMDPs) with context in hindsight. We design a novel model-base d algorithmic framework which can be instantiated with both a model-optimistic a nd a value-optimistic solver. We prove an $\tilde{O}(\sqrt{\mathsf{Var}^\star M \Gamma S A K})$ regret bound where $\tilde{O}$ hides logarithm factors, M is t he number of contexts, S is the number of states, A is the number of actions , K is the number of episodes, $\Gamma \le S$ is the maximum transition degree of any state-action pair, and $\mathsf{Var}^\star$ is a variance quantity descr ibing the determinism of the LMDP. The regret bound only scales logarithmically with the planning horizon, thus yielding the first (nearly) horizon-free regret bound for LMDP. This is also the first problem-dependent regret bound for LMDP. Key in our proof is an analysis of the total variance of alpha vectors (a genera lization of value functions), which is handled with a truncation method. We comp lement our positive result with a novel $\Omega(\sqrt{\mathsf{Var}^\star M S A K })$ regret lower bound with $\Gamma = 2$, which shows our upper bound minimax op timal when Γ is a constant for the class of variance-bounded LMDPs. Our l ower bound relies on new constructions of hard instances and an argument inspire d by the symmetrization technique from theoretical computer science, both of whi ch are technically different from existing lower bound proof for MDPs, and thus can be of independent interest.
****
Phase-aware Adversarial Defense for Improving Adversarial Robustness
Dawei Zhou, Nannan Wang, Heng Yang, Xinbo Gao, Tongliang Liu
Deep neural networks have been found to be vulnerable to adversarial noise. Rece nt works show that exploring the impact of adversarial noise on intrinsic compon ents of data can help improve adversarial robustness. However, the pattern close ly related to human perception has not been deeply studied. In this paper, inspi red by the cognitive science, we investigate the interference of adversarial noi se from the perspective of image phase, and find ordinarily-trained models lack enough robustness against phase-level perturbations. Motivated by this, we propo se a joint adversarial defense method: a phase-level adversarial training mechan ism to enhance the adversarial robustness on the phase pattern; an amplitude-bas ed pre-processing operation to mitigate the adversarial perturbation in the ampl itude pattern. Experimental results show that the proposed method can significan tly improve the robust accuracy against multiple attacks and even adaptive attac ks. In addition, ablation studies demonstrate the effectiveness of our defense s trategy.
****

From Relational Pooling to Subgraph GNNs: A Universal Framework for More Expressive Graph Neural Networks
Cai Zhou, Xiyuan Wang, Muhan Zhang
Relational pooling is a framework for building more expressive and permutation-invariant graph neural networks. However, there is limited understanding of the exact enhancement in the expressivity of RP and its connection with the Weisfeiler-Lehman hierarchy. Starting from RP, we propose to explicitly assign labels to nodes as additional features to improve graph isomorphism distinguishing power of message passing neural networks. The method is then extended to higher-dimensional WL, leading to a novel k,l-WL algorithm, a more general framework than k-WL. We further introduce the subgraph concept into our hierarchy and propose a localized k,l-WL framework, incorporating a wide range of existing work, including many subgraph GNNs. Theoretically, we analyze the expressivity of k,l-WL w.r.t. k and l and compare it with the traditional k-WL. Complexity reduction methods are also systematically discussed to build powerful and practical k,l-GNN instances. We theoretically and experimentally prove that our method is universally compatible and capable of improving the expressivity of any base GNN model. Our k,l-GNNs achieve superior performance on many synthetic and real-world datasets, which verifies the effectiveness of our framework.
****
Towards Omni-generalizable Neural Methods for Vehicle Routing Problems
Jianan Zhou, Yaoxin Wu, Wen Song, Zhiguang Cao, Jie Zhang
Learning heuristics for vehicle routing problems (VRPs) has gained much attention due to the less reliance on hand-crafted rules. However, existing methods are typically trained and tested on the same task with a fixed size and distribution (of nodes), and hence suffer from limited generalization performance. This paper studies a challenging yet realistic setting, which considers generalization across both size and distribution in VRPs. We propose a generic meta-learning framework, which enables effective training of an initialized model with the capability of fast adaptation to new tasks during inference. We further develop a simple yet efficient approximation method to reduce the training overhead. Extensive experiments on both synthetic and benchmark instances of the traveling salesman problem (TSP) and capacitated vehicle routing problem (CVRP) demonstrate the effectiveness of our method. The code is available at: https://github.com/RoyalSkye/Omni-VRP.
****
A Three-regime Model of Network Pruning
Yefan Zhou, Yaoqing Yang, Arin Chang, Michael W. Mahoney
Recent work has highlighted the complex influence training hyperparameters, e.g., the number of training epochs, can have on the prunability of machine learning models. Perhaps surprisingly, a systematic approach to predict precisely how adjusting a specific hyperparameter will affect prunability remains elusive. To address this gap, we introduce a phenomenological model grounded in the statistical mechanics of learning. Our approach uses temperature-like and load-like parameters to model the impact of neural network (NN) training hyperparameters on pruning performance. A key empirical result we identify is a sharp transition phenomenon: depending on the value of a load-like parameter in the pruned model, increasing the value of a temperature-like parameter in the pre-pruned model may either enhance or impair subsequent pruning performance. Based on this transition, we build a three-regime model by taxonomizing the global structure of the pruned NN loss landscape. Our model reveals that the dichotomous effect of high temperature is associated with transitions between distinct types of global structures in the post-pruned model. Based on our results, we present three case-studies: 1) determining whether to increase or decrease a hyperparameter for improved pruning; 2) selecting the best model to prune from a family of models; and 3) tuning the hyperparameter of the Sharpness Aware Minimization method for better pruning performance.
****
Learning to Decouple Complex Systems
Zihan Zhou, Tianshu Yu

A complex system with cluttered observations may be a coupled mixture of multiple simple sub-systems corresponding to latent entities. Such sub-systems may hold distinct dynamics in the continuous-time domain; therein, complicated interactions between sub-systems also evolve over time. This setting is fairly common in the real world but has been less considered. In this paper, we propose a sequential learning approach under this setting by decoupling a complex system for handling irregularly sampled and cluttered sequential observations. Such decoupling brings about not only subsystems describing the dynamics of each latent entity but also a meta-system capturing the interaction between entities over time. Specifically, we argue that the meta-system evolving within a simplex is governed by projected differential equations (ProjDEs). We further analyze and provide neural-friendly projection operators in the context of Bregman divergence. Experimental results on synthetic and real-world datasets show the advantages of our approach when facing complex and cluttered sequential data compared to the state-of-the-art.

****

ESC: Exploration with Soft Commonsense Constraints for Zero-shot Object Navigation

Kaiwen Zhou, Kaizhi Zheng, Connor Pryor, Yilin Shen, Hongxia Jin, Lise Getoor, Xin Eric Wang

The ability to accurately locate and navigate to a specific object is a crucial capability for embodied agents that operate in the real world and interact with objects to complete tasks. Such object navigation tasks usually require large-scale training in visual environments with labeled objects, which generalizes poorly to novel objects in unknown environments. In this work, we present a novel zero-shot object navigation method, Exploration with Soft Commonsense constraints (ESC), that transfers commonsense knowledge in pre-trained models to open-world object navigation without any navigation experience nor any other training on the visual environments. First, ESC leverages a pre-trained vision and language model for open-world prompt-based grounding and a pre-trained commonsense language model for room and object reasoning. Then ESC converts commonsense knowledge into navigation actions by modeling it as soft logic predicates for efficient exploration. Extensive experiments on MP3D, HM3D, and RoboTHOR benchmarks show that our ESC method improves significantly over baselines, and achieves new state-of-the-art results for zero-shot object navigation (e.g., 288% relative Success Rate improvement than CoW on MP3D).

****

On Strengthening and Defending Graph Reconstruction Attack with Markov Chain Approximation

Zhanke Zhou, Chenyu Zhou, Xuan Li, Jiangchao Yao, Quanming Yao, Bo Han

Although powerful graph neural networks (GNNs) have boosted numerous real-world applications, the potential privacy risk is still underexplored. To close this gap, we perform the first comprehensive study of graph reconstruction attack that aims to reconstruct the adjacency of nodes. We show that a range of factors in GNNs can lead to the surprising leakage of private links. Especially by taking GNNs as a Markov chain and attacking GNNs via a flexible chain approximation, we systematically explore the underneath principles of graph reconstruction attack, and propose two information theory-guided mechanisms: (1) the chain-based attack method with adaptive designs for extracting more private information; (2) the chain-based defense method that sharply reduces the attack fidelity with moderate accuracy loss. Such two objectives disclose a critical belief that to recover better in attack, you must extract more multi-aspect knowledge from the trained GNN; while to learn safer for defense, you must forget more link-sensitive information in training GNNs. Empirically, we achieve state-of-the-art results on six datasets and three common GNNs. The code is publicly available at: https://github.com/tmlr-group/MC-GRA.

****

Sharp Variance-Dependent Bounds in Reinforcement Learning: Best of Both Worlds in Stochastic and Deterministic Environments

Runlong Zhou, Zhang Zihan, Simon Shaolei Du

We study variance-dependent regret bounds for Markov decision processes (MDPs). Algorithms with variance-dependent regret guarantees can automatically exploit e nvironments with low variance (e.g., enjoying constant regret on deterministic M DPs). The existing algorithms are either variance-independent or suboptimal. We first propose two new environment norms to characterize the fine-grained varianc e properties of the environment. For model-based methods, we design a variant of the MVP algorithm (Zhang et al., 2021a). We apply new analysis techniques to de monstrate that this algorithm enjoys variance-dependent bounds with respect to t he norms we propose. In particular, this bound is simultaneously minimax optimal for both stochastic and deterministic MDPs, the first result of its kind. We fu rther initiate the study on model-free algorithms with variance-dependent regret bounds by designing a reference-function-based algorithm with a novel capped-do ubling reference update schedule. Lastly, we also provide lower bounds to comple ment our upper bounds.

****

Learning Unforeseen Robustness from Out-of-distribution Data Using Equivariant D omain Translator

Sicheng Zhu, Bang An, Furong Huang, Sanghyun Hong

Current approaches for training robust models are typically tailored to scenario s where data variations are accessible in the training set. While shown effectiv e in achieving robustness to these foreseen variations, these approaches are ine ffective in learning unforeseen robustness, i.e., robustness to data variations without known characterization or training examples reflecting them. In this wor k, we learn unforeseen robustness by harnessing the variations in the abundant o ut-of-distribution data. To overcome the main challenge of using such data, the domain gap, we use a domain translator to bridge it and bound the unforeseen rob ustness on the target distribution. As implied by our analysis, we propose a two -step algorithm that first trains an equivariant domain translator to map out-of -distribution data to the target distribution while preserving the considered va riation, and then regularizes a model’s output consistency on the domain-transla ted data to improve its robustness. We empirically show the effectiveness of our approach in improving unforeseen and foreseen robustness compared to existing a pproaches. Additionally, we show that training the equivariant domain translator serves as an effective criterion for source data selection.

****

Markovian Gaussian Process Variational Autoencoders

Harrison Zhu, Carles Balsells-Rodas, Yingzhen Li

Sequential VAEs have been successfully considered for many high-dimensional time series modelling problems, with many variant models relying on discrete-time me chanisms such as recurrent neural networks (RNNs). On the other hand, continuous -time methods have recently gained attraction, especially in the context of irre gularly-sampled time series, where they can better handle the data than discrete -time methods. One such class are Gaussian process variational autoencoders (GPV AEs), where the VAE prior is set as a Gaussian process (GP). However, a major li mitation of GPVAEs is that it inherits the cubic computational cost as GPs, maki ng it unattractive to practioners. In this work, we leverage the equivalent disc rete state space representation of Markovian GPs to enable linear time GPVAE tra ining via Kalman filtering and smoothing. For our model, Markovian GPVAE (MGPVAE), we show on a variety of high-dimensional temporal and spatiotemporal tasks th at our method performs favourably compared to existing approaches whilst being c omputationally highly scalable.

****

Mixture Proportion Estimation Beyond Irreducibility

Yilun Zhu, Aaron Fjeldsted, Darren Holland, George Landon, Azaree Lintereur, Cla yton Scott

The task of mixture proportion estimation (MPE) is to estimate the weight of a c omponent distribution in a mixture, given observations from both the component a nd mixture. Previous work on MPE adopts the irreducibility assumption, which ens ures identifiablity of the mixture proportion. In this paper, we propose a more general sufficient condition that accommodates several settings of interest wher

e irreducibility does not hold. We further present a resampling-based meta-algorithm that takes any existing MPE algorithm designed to work under irreducibility and adapts it to work under our more general condition. Our approach empirically exhibits improved estimation performance relative to baseline methods and to a recently proposed regrouping-based algorithm.
****
Exploring Model Dynamics for Accumulative Poisoning Discovery
Jianing Zhu, Xiawei Guo, Jiangchao Yao, Chao Du, Li He, Shuo Yuan, Tongliang Liu, Liang Wang, Bo Han
Adversarial poisoning attacks pose huge threats to various machine learning applications. Especially, the recent accumulative poisoning attacks show that it is possible to achieve irreparable harm on models via a sequence of imperceptible attacks followed by a trigger batch. Due to the limited data-level discrepancy in real-time data streaming, current defensive methods are indiscriminate in handling the poison and clean samples. In this paper, we dive into the perspective of model dynamics and propose a novel information measure, namely, Memorization Discrepancy, to explore the defense via the model-level information. By implicitly transferring the changes in the data manipulation to that in the model outputs, Memorization Discrepancy can discover the imperceptible poison samples based on their distinct dynamics from the clean samples. We thoroughly explore its properties and propose Discrepancy-aware Sample Correction (DSC) to defend against accumulative poisoning attacks. Extensive experiments comprehensively characterized Memorization Discrepancy and verified its effectiveness. The code is publicly available at: https://github.com/tmlr-group/Memorization-Discrepancy.
****
Decentralized SGD and Average-direction SAM are Asymptotically Equivalent
Tongtian Zhu, Fengxiang He, Kaixuan Chen, Mingli Song, Dacheng Tao
Decentralized stochastic gradient descent (D-SGD) allows collaborative learning on massive devices simultaneously without the control of a central server. However, existing theories claim that decentralization invariably undermines generalization. In this paper, we challenge the conventional belief and present a completely new perspective for understanding decentralized learning. We prove that D-SGD implicitly minimizes the loss function of an average-direction Sharpness-aware minimization (SAM) algorithm under general non-convex non-β-smooth settings. This surprising asymptotic equivalence reveals an intrinsic regularization-optimization trade-off and three advantages of decentralization: (1) there exists a free uncertainty evaluation mechanism in D-SGD to improve posterior estimation; (2) D-SGD exhibits a gradient smoothing effect; and (3) the sharpness regularization effect of D-SGD does not decrease as total batch size increases, which justifies the potential generalization benefit of D-SGD over centralized SGD (C-SGD) in large-batch scenarios.
****
Principled Reinforcement Learning with Human Feedback from Pairwise or K-wise Comparisons
Banghua Zhu, Michael Jordan, Jiantao Jiao
We provide a theoretical framework for Reinforcement Learning with Human Feedback (RLHF). We show that when the underlying true reward is linear, under both Bradley-Terry-Luce (BTL) model (pairwise comparison) and Plackett-Luce (PL) model (K-wise comparison), MLE converges under certain semi-norm for the family of linear reward. On the other hand, when training a policy based on the learned reward model, we show that MLE fails while a pessimistic MLE provides policies with good performance under certain coverage assumption. We also show that under the PL model, both the true MLE and a different MLE which splits the K-wise comparison into pairwise comparisons converge, while the true MLE is asymptotically more efficient. Our results validate the empirical success of the existing RLHF algorithms, and provide new insights for algorithm design. Our analysis can also be applied for the problem of online RLHF and inverse reinforcement learning.
****
Unleashing Mask: Explore the Intrinsic Out-of-Distribution Detection Capability
Jianing Zhu, Hengzhuang Li, Jiangchao Yao, Tongliang Liu, Jianliang Xu, Bo Han

Out-of-distribution (OOD) detection is an indispensable aspect of secure AI when deploying machine learning models in real-world applications. Previous paradigms either explore better scoring functions or utilize the knowledge of outliers to equip the models with the ability of OOD detection. However, few of them pay attention to the intrinsic OOD detection capability of the given model. In this work, we generally discover the existence of an intermediate stage of a model trained on in-distribution (ID) data having higher OOD detection performance than that of its final stage across different settings, and further identify one critical data-level attribution to be learning with the atypical samples. Based on such insights, we propose a novel method, Unleashing Mask, which aims to restore the OOD discriminative capabilities of the well-trained model with ID data. Our method utilizes a mask to figure out the memorized atypical samples, and then finetune the model or prune it with the introduced mask to forget them. Extensive experiments and analysis demonstrate the effectiveness of our method. The code is available at: https://github.com/tmlr-group/Unleashing-Mask.

****

Benign Overfitting in Deep Neural Networks under Lazy Training

Zhenyu Zhu, Fanghui Liu, Grigorios Chrysos, Francesco Locatello, Volkan Cevher

This paper focuses on over-parameterized deep neural networks (DNNs) with ReLU activation functions and proves that when the data distribution is well-separated, DNNs can achieve Bayes-optimal test error for classification while obtaining (nearly) zero-training error under the lazy training regime. For this purpose, we unify three interrelated concepts of overparameterization, benign overfitting, and the Lipschitz constant of DNNs. Our results indicate that interpolating with smoother functions leads to better generalization. Furthermore, we investigate the special case where interpolating smooth ground-truth functions is performed by DNNs under the Neural Tangent Kernel (NTK) regime for generalization. Our result demonstrates that the generalization error converges to a constant order that only depends on label noise and initialization noise, which theoretically verifies benign overfitting. Our analysis provides a tight lower bound on the normalized margin under non-smooth activation functions, as well as the minimum eigenvalue of NTK under high-dimensional settings, which has its own interest in learning theory.

****

Interpolation for Robust Learning: Data Augmentation on Wasserstein Geodesics

Jiacheng Zhu, Jielin Qiu, Aritra Guha, Zhuolin Yang, Xuanlong Nguyen, Bo Li, Ding Zhao

We propose to study and promote the robustness of a model as per its performance on a continuous geodesic interpolation of subpopulations, e.g., a class of samples in a classification problem. Specifically, (1) we augment the data by finding the worst-case Wasserstein barycenter on the geodesic connecting subpopulation distributions. (2) we regularize the model for smoother performance on the continuous geodesic path connecting subpopulation distributions. (3) Additionally, we provide a theoretical guarantee of robustness improvement and investigate how the geodesic location and the sample size contribute, respectively. Experimental validations of the proposed strategy on four datasets including CIFAR-100 and ImageNet, establish the efficacy of our method, e.g., our method improves the baselines' certifiable robustness on CIFAR10 upto 7.7%, with 16.8% on empirical robustness on CIFAR-100. Our work provides a new perspective of model robustness through the lens of Wasserstein geodesic-based interpolation with a practical off-the-shelf strategy that can be combined with existing robust training methods.

****

LeadFL: Client Self-Defense against Model Poisoning in Federated Learning

Chaoyi Zhu, Stefanie Roos, Lydia Y. Chen

Federated Learning is highly susceptible to backdoor and targeted attacks as participants can manipulate their data and models locally without any oversight on whether they follow the correct process. There are a number of server-side defenses that mitigate the attacks by modifying or rejecting local updates submitted by clients. However, we find that bursty adversarial patterns with a high variance in the number of malicious clients can circumvent the existing defenses. We p

ropose a client-self defense, LeadFL, that is combined with existing server-side defenses to thwart backdoor and targeted attacks. The core idea of LeadFL is a novel regularization term in local model training such that the Hessian matrix of local gradients is nullified. We provide the convergence analysis of LeadFL and its robustness guarantee in terms of certified radius. Our empirical evaluation shows that LeadFL is able to mitigate bursty adversarial patterns for both iid and non-iid data distributions. It frequently reduces the backdoor accuracy from more than 75% for state-of-the-art defenses to less than 10% while its impact on the main task accuracy is always less than for other client-side defenses.

****

XTab: Cross-table Pretraining for Tabular Transformers

Bingzhao Zhu, Xingjian Shi, Nick Erickson, Mu Li, George Karypis, Mahsa Shoaran

The success of self-supervised learning in computer vision and natural language processing has motivated pretraining methods on tabular data. However, most existing tabular self-supervised learning models fail to leverage information across multiple data tables and cannot generalize to new tables. In this work, we introduce XTab, a framework for cross-table pretraining of tabular transformers on datasets from various domains. We address the challenge of inconsistent column types and quantities among tables by utilizing independent featurizers and using federated learning to pretrain the shared component. Tested on 84 tabular prediction tasks from the OpenML-AutoML Benchmark (AMLB), we show that (1) XTab consistently boosts the generalizability, learning speed, and performance of multiple tabular transformers, (2) by pretraining FT-Transformer via XTab, we achieve superior performance than other state-of-the-art tabular deep learning models on various tasks such as regression, binary, and multiclass classification.

****

Provable Multi-instance Deep AUC Maximization with Stochastic Pooling

Dixian Zhu, Bokun Wang, Zhi Chen, Yaxing Wang, Milan Sonka, Xiaodong Wu, Tianbao Yang

This paper considers a novel application of deep AUC maximization (DAM) for multi-instance learning (MIL), in which a single class label is assigned to a bag of instances (e.g., multiple 2D slices of a CT scan for a patient). We address a neglected yet non-negligible computational challenge of MIL in the context of DAM, i.e., bag size is too large to be loaded into GPU memory for backpropagation, which is required by the standard pooling methods of MIL. To tackle this challenge, we propose variance-reduced stochastic pooling methods in the spirit of stochastic optimization by formulating the loss function over the pooled prediction as a multi-level compositional function. By synthesizing techniques from stochastic compositional optimization and non-convex min-max optimization, we propose a unified and provable muli-instance DAM (MIDAM) algorithm with stochastic smoothed-max pooling or stochastic attention-based pooling, which only samples a few instances for each bag to compute a stochastic gradient estimator and to update the model parameter. We establish a similar convergence rate of the proposed MIDAM algorithm as the state-of-the-art DAM algorithms. Our extensive experiments on conventional MIL datasets and medical datasets demonstrate the superiority of our MIDAM algorithm. The method is open-sourced at https://libauc.org/.

****

Surrogate Model Extension (SME): A Fast and Accurate Weight Update Attack on Federated Learning

Junyi Zhu, Ruicong Yao, Matthew B. Blaschko

In Federated Learning (FL) and many other distributed training frameworks, collaborators can hold their private data locally and only share the network weights trained with the local data after multiple iterations. Gradient inversion is a family of privacy attacks that recovers data from its generated gradients. Seemingly, FL can provide a degree of protection against gradient inversion attacks on weight updates, since the gradient of a single step is concealed by the accumulation of gradients over multiple local iterations. In this work, we propose a principled way to extend gradient inversion attacks to weight updates in FL, thereby better exposing weaknesses in the presumed privacy protection inherent in FL. In particular, we propose a surrogate model method based on the characteristic

of two-dimensional gradient flow and low-rank property of local updates. Our method largely boosts the ability of gradient inversion attacks on weight updates containing many iterations and achieves state-of-the-art (SOTA) performance. Additionally, our method runs up to $100\times$ faster than the SOTA baseline in the common FL scenario. Our work re-evaluates and highlights the privacy risk of sharing network weights. Our code is available at https://github.com/JunyiZhu-AI/surrogate_model_extension.
****
Weak Proxies are Sufficient and Preferable for Fairness with Missing Sensitive Attributes
Zhaowei Zhu, Yuanshun Yao, Jiankai Sun, Hang Li, Yang Liu
Evaluating fairness can be challenging in practice because the sensitive attributes of data are often inaccessible due to privacy constraints. The go-to approach that the industry frequently adopts is using off-the-shelf proxy models to predict the missing sensitive attributes, e.g. Meta (Alao et al., 2021) and Twitter (Belli et al., 2022). Despite its popularity, there are three important questions unanswered: (1) Is directly using proxies efficacious in measuring fairness? (2) If not, is it possible to accurately evaluate fairness using proxies only? (3) Given the ethical controversy over infer-ring user private information, is it possible to only use weak (i.e. inaccurate) proxies in order to protect privacy? Our theoretical analyses show that directly using proxy models can give a false sense of (un)fairness. Second, we develop an algorithm that is able to measure fairness (provably) accurately with only three properly identified proxies. Third, we show that our algorithm allows the use of only weak proxies (e.g. with only 68.85% accuracy on COMPAS), adding an extra layer of protection on user privacy. Experiments validate our theoretical analyses and show our algorithm can effectively measure and mitigate bias. Our results imply a set of practical guidelines for prac-titioners on how to use proxies properly. Code is available at https://github.com/UCSC-REAL/fair-eval.
****
Label Distributionally Robust Losses for Multi-class Classification: Consistency, Robustness and Adaptivity
Dixian Zhu, Yiming Ying, Tianbao Yang
We study a family of loss functions named label-distributionally robust (LDR) losses for multi-class classification that are formulated from distributionally robust optimization (DRO) perspective, where the uncertainty in the given label information are modeled and captured by taking the worse case of distributional weights. The benefits of this perspective are several fold: (i) it provides a unified framework to explain the classical cross-entropy (CE) loss and SVM loss and their variants, (ii) it includes a special family corresponding to the temperature-scaled CE loss, which is widely adopted but poorly understood; (iii) it allows us to achieve adaptivity to the uncertainty degree of label information at an instance level. Our contributions include: (1) we study both consistency and robustness by establishing top-k ($\forall k\geq 1$) consistency of LDR losses for multi-class classification, and a negative result that a top-1 consistent and symmetric robust loss cannot achieve top-k consistency simultaneously for all $k\geq 2$; (2) we propose a new adaptive LDR loss that automatically adapts the individualized temperature parameter to the noise degree of class label of each instance; (3) we demonstrate stable and competitive performance for the proposed adaptive LDR loss on 7 benchmark datasets under 6 noisy label and 1 clean settings against 13 loss functions, and on one real-world noisy dataset. The method is open-sourced at https://github.com/Optimization-AI/ICML2023_LDR.
****
Likelihood Adjusted Semidefinite Programs for Clustering Heterogeneous Data
Yubo Zhuang, Xiaohui Chen, Yun Yang
Clustering is a widely deployed unsupervised learning tool. Model-based clustering is a flexible framework to tackle data heterogeneity when the clusters have different shapes. Likelihood-based inference for mixture distributions often involves non-convex and high-dimensional objective functions, imposing difficult computational and statistical challenges. The classic expectation-maximization (EM)

algorithm is a computationally thrifty iterative method that maximizes a surrogate function minorizing the log-likelihood of observed data in each iteration, which however suffers from bad local maxima even in the special case of the standard Gaussian mixture model with common isotropic covariance matrices. On the other hand, recent studies reveal that the unique global solution of a semidefinite programming (SDP) relaxed K-means achieves the information-theoretically sharp threshold for perfectly recovering the cluster labels under the standard Gaussian mixture model. In this paper, we extend the SDP approach to a general setting by integrating cluster labels as model parameters and propose an iterative likelihood adjusted SDP (iLA-SDP) method that directly maximizes the exact observed likelihood in the presence of data heterogeneity. By lifting the cluster assignment to group-specific membership matrices, iLA-SDP avoids centroids estimation – a key feature that allows exact recovery under well-separateness of centroids without being trapped by their adversarial configurations. Thus iLA-SDP is less sensitive than EM to initialization and more stable on high-dimensional data. Our numeric experiments demonstrate that iLA-SDP can achieve lower mis-clustering errors over several widely used clustering methods including K-means, SDP and EM algorithms.

****

Are Random Decompositions all we need in High Dimensional Bayesian Optimisation?

Juliusz Krzysztof Ziomek, Haitham Bou Ammar

Learning decompositions of expensive-to-evaluate black-box functions promises to scale Bayesian optimisation (BO) to high-dimensional problems. However, the success of these techniques depends on finding proper decompositions that accurately represent the black-box. While previous works learn those decompositions based on data, we investigate data-independent decomposition sampling rules in this paper. We find that data-driven learners of decompositions can be easily misled towards local decompositions that do not hold globally across the search space. Then, we formally show that a random tree-based decomposition sampler exhibits favourable theoretical guarantees that effectively trade off maximal information gain and functional mismatch between the actual black-box and its surrogate as provided by the decomposition. Those results motivate the development of the random decomposition upper-confidence bound algorithm (RDUCB) that is straightforward to implement - (almost) plug-and-play - and, surprisingly, yields significant empirical gains compared to the previous state-of-the-art on a comprehensive set of benchmarks. We also confirm the plug-and-play nature of our modelling component by integrating our method with HEBO, showing improved practical gains in the highest dimensional tasks from Bayesmark problem suite.

****

Revisiting Bellman Errors for Offline Model Selection

Joshua P Zitovsky, Daniel De Marchi, Rishabh Agarwal, Michael Rene Kosorok

Offline model selection (OMS), that is, choosing the best policy from a set of many policies given only logged data, is crucial for applying offline RL in real-world settings. One idea that has been extensively explored is to select policies based on the mean squared Bellman error (MSBE) of the associated Q-functions. However, previous work has struggled to obtain adequate OMS performance with Bellman errors, leading many researchers to abandon the idea. To this end, we elucidate why previous work has seen pessimistic results with Bellman errors and identify conditions under which OMS algorithms based on Bellman errors will perform well. Moreover, we develop a new estimator of the MSBE that is more accurate than prior methods. Our estimator obtains impressive OMS performance on diverse discrete control tasks, including Atari games.

****

spred: Solving L1 Penalty with SGD

Liu Ziyin, Zihao Wang

We propose to minimize a generic differentiable objective with L_1 constraint using a simple reparametrization and straightforward stochastic gradient descent. Our proposal is the direct generalization of previous ideas that the L_1 penalty may be equivalent to a differentiable reparametrization with weight decay. We prove that the proposed method, spred, is an exact differentiable solver of $

L_1$ and that the reparametrization trick is completely “benign" for a generic n
onconvex function. Practically, we demonstrate the usefulness of the method in (
1) training sparse neural networks to perform gene selection tasks, which involv
es finding relevant features in a very high dimensional space, and (2) neural ne
twork compression task, to which previous attempts at applying the L_1-penalty
 have been unsuccessful. Conceptually, our result bridges the gap between the sp
arsity in deep learning and conventional statistical learning.
****

The Benefits of Mixup for Feature Learning

Difan Zou, Yuan Cao, Yuanzhi Li, Quanquan Gu

Mixup, a simple data augmentation method that randomly mixes two data points via
 linear interpolation, has been extensively applied in various deep learning app
lications to gain better generalization. However, its theoretical explanation re
mains largely unclear. In this work, we aim to seek a fundamental understanding
of the benefits of Mixup. We first show that Mixup using different linear interp
olation parameters for features and labels can still achieve similar performance
 as standard Mixup. This indicates that the intuitive linearity explanation in Z
hang et al., (2018) may not fully explain the success of Mixup. Then, we perform
 a theoretical study of Mixup from the feature learning perspective. We consider
 a feature-noise data model and show that Mixup training can effectively learn t
he rare features (appearing in a small fraction of data) from its mixture with t
he common features (appearing in a large fraction of data). In contrast, standar
d training can only learn the common features but fails to learn the rare featur
es, thus suffering from bad generalization performance. Moreover, our theoretica
l analysis also shows that the benefits of Mixup for feature learning are mostly
 gained in the early training phase, based on which we propose to apply early st
opping in Mixup. Experimental results verify our theoretical findings and demons
trate the effectiveness of the early-stopped Mixup training.
****